This draft evidence report/technology assessment is distributed solely for the purpose of pre-release peer review. It has not been otherwise disseminated by AHRQ. It does not represent, and should not be construed to represent, an AHRQ determination or policy.

Draft Methods Research Report

Number XX

Findings of Bayesian Mixed Treatment Comparison Meta-Analyses: Comparison and Exploration Using Real-World Data and Simulation

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. xxx-xx-xxxx

Prepared by:
<EPC> Evidence-based Practice Center
<City, State>

Investigators:

AHRQ Publication No. xx-EHCxxx
<Month Year>

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted for which further reproduction is prohibited without the specific permission of copyright holders.

Suggested Citation:

<Authors> Findings of Bayesian Mixed Treatment Comparison Meta-Analyses: Comparison and Exploration Using Real-World Data and Simulation. <EPC> Evidence-based Practice Center, Contract No. <Contract Number>. (AHRQ to provide the remainder)

This information is distributed solely for the purposes of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the Agency for Healthcare Research and Quality. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services determination or policy.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-Based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome comments on this evidence report. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Parivash Nourjah, Ph.D.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: <Acknowledgments>.

Peer Reviewers

<Name>
<Place>
<City>, <ST>

<Name>
<Place>
<City>, <ST>

Findings of Bayesian Mixed Treatment Comparison Meta-Analyses: Comparison and Exploration Using Real-World Data and Simulation

Structured Abstract

Objectives: The overarching aim of this research was to better understand the performance of Bayesian MTC methods. Specific objectives were to examine (1a) how results of Bayesian MTC methods compare with several more commonly considered frequentist indirect methods, (1b) how Bayesian MTC methods perform for different evidence network patterns, (2) how meta-regression can be used with Bayesian MTC meta-analysis to explore heterogeneity, and (3) how findings of Bayesian MTC meta-analyses compare for different numbers of studies and different network pattern assumptions.

Methods: For objectives 1 and 2, we used data from two recent CERs; one examined second-generation antidepressants (SGAs) and one examined biologic disease modifying antirheumatic drugs (DMARDs) for rheumatoid arthritis (RA). For objective 1, we compared results of Bayesian MTC methods with those of three frequentist indirect methods: meta-regression, the Bucher method, and logistic regression for dichotomous and continuous outcomes. For objective 2, we conducted two types of meta-regression. One explored subgroup effects with a binary covariate to assess whether efficacy of SGAs differs between older adults (≥55 years) and adults of any age. The other explored a continuous covariate to assess whether treatment efficacy varies by disease duration of RA. For objective 3, we used simulated data to examine the Bayesian MTC method's ability to produce valid results for two data scenarios when varying numbers of studies were available for each comparison for various network patterns.

Results: Bayesian MTC methods permitted the calculation of results for more comparisons of interest than frequentist meta-regression or the Bucher method (when applied as they would typically be used). When comparisons were calculated, the findings generally agreed, but differed for a small proportion (less than 10%) of comparisons. Regarding precision, logistic regression produced the most precise estimates, followed by the Bayesian MTC method.

Our meta-regressions found a trend toward lesser efficacy for SGAs in older adults and a trend toward greater efficacy of biologic DMARDs for those with greater mean disease duration.

Our simulations supported the validity of Bayesian MTC methods for star and ladder network patterns but raised some concerns about one closed loop (and possibly loop) network patterns. Simulations generally found similar results for scenarios when only one study was available for each comparison and those when more studies (two, three, five, or ten) were available.

Conclusions: Bayesian MTC methods offer several advantages over frequentist indirect methods, including the ability to produce results for all comparisons of interest in a single analysis. Results of Bayesian MTC methods and those of frequentist indirect methods may differ for a small proportion of comparisons, which could lead to differences in conclusions when using different methods. Our findings raise some concerns about the validity of the results of Bayesian MTC methods for certain network patterns. Further research is needed to explore

additional real-world datasets and simulated data to determine if our findings are generalizable and to better understand the validity of Bayesian MTC methods for various scenarios.

Contents

Introduction 11
Background 11
Evidence Networks 12
Second-Generation Antidepressants 12
Biologic Disease Modifying Antirheumatic Drugs for Rheumatoid Arthritis 14
Scope and Key Questions 16
Methods 19
Data Included in this Report 19
Dataset 1: Second-Generation Antidepressants 19
Dataset 2: Biologic Disease Modifying Antirheumatic Drugs for Treatment of Rheumatoid Arthritis 19
Key Question 1: Comparison of Bayesian MTC Meta-Analysis with Frequentist Indirect Methods 20
Bayesian Mixed Treatment Comparison Meta-analysis 21
Frequentist Indirect Comparisons 23
Choice of Network Patterns 26
Key Question 2: Meta-Regression 26
Meta-Regression With a Subgroup Indicator Covariate: Efficacy of SGAs and Older Adults 27
Meta-Regression With a Continuous Covariate: Efficacy of Biologic DMARDs and Disease Duration of Rheumatoid Arthritis 27
Key Question 3: Methods Section for Simulation Study 28
Results 31
Key Question 1. Bayesian Mixed Treatment Comparison (MTC) Methods Compared With Frequentist Indirect Methods, and Performance for Different Types of Evidence Network Patterns 31
Comparison of Bayesian MTC Results with those of Various Frequentist Analyses for the Full Networks 31
Second-generation Antidepressants: Response (Binary Outcome) 31
Second-generation Antidepressants: Mean Change in HAM-D (Continuous Outcome) 32
Biologic Disease Modifying Antirheumatic Drugs for Rheumatoid Arthritis: ACR50 (Binary Outcome) 33
Performance for Different Types of Evidence Network Patterns 34
Comparison of Bayesian MTC Results for the Full Networks with Bayesian MTC Results for Sub-components 38
Key Question 2. Meta-Regression with Bayesian MTC Meta-Analysis 44
Meta-Regression with a Subgroup Indicator Covariate 45
Meta-regression with a Continuous Covariate 47

Key Question 3. Stability of the Bayesian MTC Analyses for Different Numbers of Studies and Network Pattern Assumptions 49
Equivalent Efficacy Scenario 50
Superior Efficacy Scenario 51
Discussion 55
Main Findings by Key Question 55
Key Question 1: Bayesian Mixed Treatment Comparison (MTC) Methods Compared With Several Frequentist Indirect Methods, and Performance for Different Types of Evidence Network Patterns 55
Key Question 2: Meta-Regression with Bayesian MTC Meta-Analysis 56
Key Question 3: Stability of the Bayesian MTC Analyses for Different Numbers of Studies and Network Pattern Assumptions 57
Implications for Comparative Effectiveness Reviews 58
Limitations 59
Future Research 60
Conclusions 60
References 62

Tables

Table 1. Second-generation antidepressants approved for use in the United States 13
Table 2. Biologic DMARD treatments for rheumatoid arthritis 15
Table 3. Comparison of findings for SGA response between Bayesian MTC meta-analysis findings and those of frequentist methods: agreement and kappas 32
Table 4. Comparison of findings for mean change in HAM-D between Bayesian MTC meta-analysis findings and those of frequentist methods: agreement and kappas 33
Table 5. Comparison of findings for ACR50 between Bayesian MTC meta-analysis findings and those of frequentist methods: agreement and kappas 34
Table 6. Comparison of findings for SGA response between Bayesian MTC meta-analysis findings and those of frequentist methods for various network patterns 35
Table 7. Comparison of findings for mean change in HAM-D between Bayesian MTC meta-analysis findings and those of frequentist methods for various network patterns 36
Table 8. Comparison of findings for ACR50 response between Bayesian MTC meta-analysis findings and those of frequentist methods for various network patterns 37
Table 9. Measures of model fit for second-generation antidepressants for response 38
Table 10. Probability of best treatment for second-generation antidepressants for response 39
Table 11. Measures of model fit for second-generation antidepressants for mean change from baseline in HAM-D 41
Table 12. Measures of model fit for biologic DMARDs for rheumatoid arthritis for achieving ACR50 43

Table 13. Probability of best treatment for biologic DMARDs for rheumatoid arthritis for achieving ACR50 43
Table 14. Measures of model fit between main analysis and meta-regression 47
Table 15. Interaction effect and probability of best treatment 48
Table 16. Simulation results: Probability of best treatment under equivalent efficacy scenario 50
Table 17. Bias of Simulation Results: Difference between observed and expected probability of best treatment under equivalent efficacy scenario 51
Table 18. Simulation results: Probability of best treatment under superior efficacy scenario 52
Table 19. Bias of simulation results: Difference between observed and expected probability of best treatment under superior efficacy scenario 53
Table 20. Advantages and disadvantages of the Bayesian MTC approach 58
Table 21. Possible targets for future research, by Key Question (KQ) 61

Figures

Figure 1. Evidence network patterns 12
Figure 2. Evidence network for mixed treatment comparison meta-analysis of second-generation antidepressants for achieving response for adults with depression 14
Figure 3. Evidence network for mixed treatment comparison meta-analysis of biologic DMARDs for achieving ACR 50 for adults with rheumatoid arthritis 16
Figure 4. Simulation study design 29
Figure 5. Results of Bayesian MTC meta-analysis for five scenarios: odds ratio (95% credible interval) comparing fluoxetine, paroxetine, and sertraline for achieving response 40
Figure 6. Results of Bayesian MTC meta-analysis for five scenarios: weighted mean difference (95% credible interval) comparing escitalopram, fluoxetine, and venlafaxine for mean change from baseline in HAM-D 42
Figure 7. Results of Bayesian MTC meta-analysis for five scenarios: odds ratio (95% credible interval) comparing abatacept and infliximab for treatment response (ACR50) 44
Figure 8. Evidence network for subgroup meta-regression to assess whether efficacy of second-generation antidepressants differs for older adults 45
Figure 9. Odds ratios (95% credible interval) of treatment response for second-generation antidepressants, by patient population 46
Figure 10. Evidence network for continuous covariate meta-regression to assess whether efficacy of biologic DMARDs differs for patients with longer disease duration 47
Figure 11. Odds ratios (95% credible interval) of treatment response for Etanercept compared with other biologic DMARDs, with and without disease duration covariate 49

Figure 12. One closed loop network pattern used in simulation 58

Appendices

Appendix A. WinBUGS code used in Bayesian Mixed Treatment Comparisons Meta-Analysis A-1

Appendix B. Summary of Studies Included in Analyses B-1

Appendix C. Evidence Networks C-1

Appendix D. Results of Analyses D-1

Appendix E. Comparison of Precision E-1

Appendix F. Distribution Plots from Simulation Study F-1

Introduction

Background

Comparative effectiveness reviews (CERs) often aim to compare the benefits and harms of multiple available approaches for treating a health condition[1] with the ultimate goal of informing clinical practice and other decisionmaking. To this end, analysts conducting CERs aim to find studies conducting direct head-to-head comparisons. However, direct head-to-head evidence on competing interventions is often scant. Therefore, several methods to conduct indirect comparisons have been proposed.[2-7] These include meta-regression, logistic regression, adjusted indirect comparisons using a frequentist approach,[2-4] and, more recently, Bayesian mixed treatment comparison (MTC) meta-analysis.[5-7]

MTC meta-analysis is a relatively new methodology.[8] Various other terms have been used to describe the approach, including multiple treatment comparisons[9-11] and network meta-analysis.[12, 13] One of the most compelling reasons to use MTC meta-analysis is that it allows for the combination of both direct head-to-head and indirect evidence (e.g., placebo-controlled trials) into one modeling framework. The use of all potentially relevant available evidence is an appealing feature for analysts, since other methods rely solely on one type of evidence. In addition, unlike other indirect analysis methods, MTC meta-analysis allows all relevant comparisons to be made through a single analysis, providing the information to calculate an effect size for each comparison of interest and to rank treatments based on the probability of being the best treatment.

The history of MTC meta-analysis dates back to 1996 when Higgins and Whitehead first described likelihood-based methods for indirect comparisons, focusing on Bayesian methods and providing an illustration of the methodology using data from 26 clinical trials that investigated the prevention of cirrhosis using beta-blockers and sclerotherapy.[14] Lu and Ades subsequently published additional information on the theoretical underpinnings of the MTC method,[8] but more research is needed to better understand how these methods operate in real-world scenarios. As investigators conducting systematic reviews use these methods and continue to develop new techniques based on these methods, it is important for us to have a better understanding of how these methods compare with other indirect methods and how MTC meta-analyses perform in various situations.

Relatively little information on the validity of MTC meta-analysis exists (in comparison with other indirect methods or for various types of evidence networks), and further research is needed.[12, 15] Some analysts have validated frequentist approaches for indirect comparisons using artificial/simulated data.[3]

One underlying question is whether the extra complexity of Bayesian MTC meta-analyses is worth the investment. MTC meta-analysis requires greater statistical expertise and can currently only be run using programs (e.g., WinBUGS) that are unfamiliar to many analysts. In contrast, several of the frequentist methods for indirect comparisons can be conducted by analysts with less statistical expertise and using programs that are more familiar (e.g., STATA). It is important to gain a better understanding of the consequences of choosing various analytic approaches. Some questions include the following: Do results differ for various analytic approaches? Do certain methods yield more precise estimates? Do the findings and validity of various analytic approaches vary depending on the evidence network pattern?

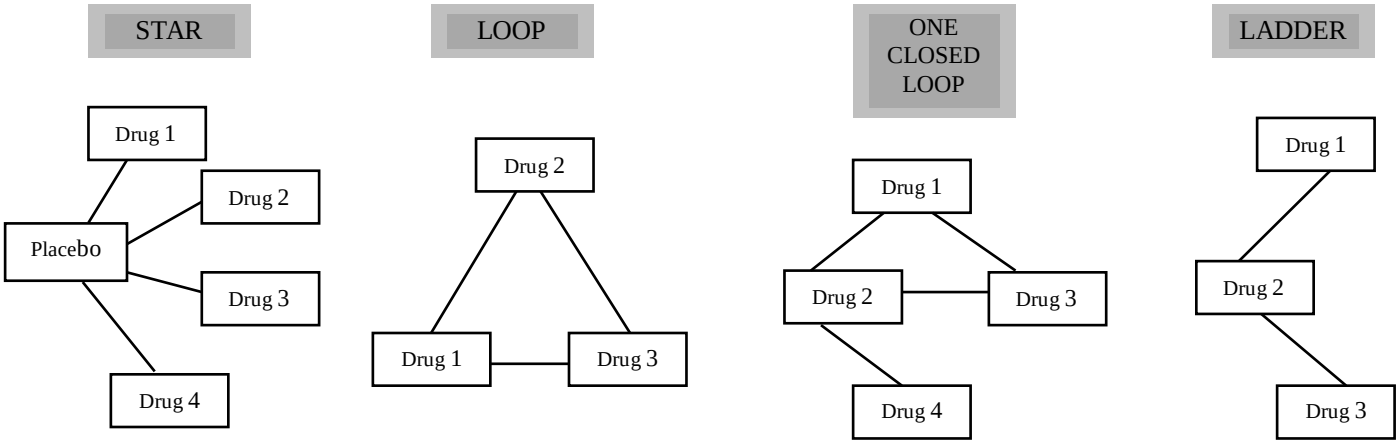
We used MTC meta-analyses in two recent CERs for the Agency for Healthcare Research and Quality (AHRQ)—one on second-generation antidepressants (SGAs)[16] and one on

treatments for rheumatoid arthritis (RA).[17] In this report, we use the real-world literature from these two reports to address our objectives and Key Questions (below).

Evidence Networks

An evidence network refers to the linkage of treatment comparisons that exists in the literature for a given population. It can take on many shapes or patterns.[18, 19] The basic premise underlying MTC methods is that the network must be a connected one. A network can include any of the four patterns below (Figure 1). A network pattern can resemble a star, with one common comparator at the center and other treatments connected through this comparator. This is a common scenario for pharmacotherapies, as RCTs often include only drug comparisons with placebo (and no direct head-to-head comparisons). Another pattern is a loop design, where all drugs or treatments are connected to one another through one other treatment. A third common pattern is a variation on this loop design, but this network also includes one or more drugs or treatments outside of the loop and is referred to as one closed loop. Another network pattern resembles a ladder, with no treatment compared to any other treatment more than once.

Figure 1. Evidence network patterns



Second-Generation Antidepressants

Our report on SGAs compared the benefits and harms of 13 SGAs approved for use in the United States[16] (Table 1).

Table 1. Second-generation antidepressants approved for use in the United States

Generic Name	U.S. Trade Name[a]	Dosage Forms	Therapeutic Classification
Bupropion[b]	Wellbutrin®; Wellbutrin SR®; Wellbutrin XL®	75, 100 mg tabs; 100, 150, 200 mg SR tabs 150, 300 mg XL tabs	Other
Citalopram[b]	Celexa®	10, 20, 40 mg tabs; 2 mg/ml solution	SSRI
Desvenlafaxine	Pristiq®	50, 100 mg tabs	SNRI
Duloxetine	Cymbalta®	20, 30, 60 mg caps	SSNRI
Escitalopram	Lexapro®	5, 10, 20 mg tabs 1 mg/ml solution	SSRI
Fluoxetine[b]	Prozac®; Prozac Weekly®	10, 20, 40 mg caps; 4mg/ml solution 90 mg caps	SSRI
Fluvoxamine[b]	Luvox®	25, 50, 100 mg tabs	SSRI
Mirtazapine[b]	Remer on® Remer on Sol tab®	15, 30, 45 mg tabs; 15, 30, 45 mg orally disintegrating tabs	SNRI[c]
Nefazodone[b]	Serzone®[d]	50, 100, 150, 200, 250 mg tabs	Other
Paroxetine[b]	Paxil®; Paxil CR®	10, 20, 30, 40 mg tabs; 2 mg/ml solution; 12.5, 25, 37.5 mg CR tabs	SSRI
Sertraline[b]	Zoloft®	25, 50, 100 mg tabs; 20 mg/ml solution	SSRI
Trazodone[b]	Desyrel®	50, 100, 150, 300 mg tabs	Other
Venlafaxine[b]	Effexor®; Effexor XR®	25, 37.5, 50, 75, 100 mg tabs; 37.5, 75, 150 mg XR caps	SNRI

[a]CR, SR, XL, and XR are registered trademarks referring to controlled, sustained, or extended-release dosage forms, respectively.

[b]Generic available for some dosage forms.

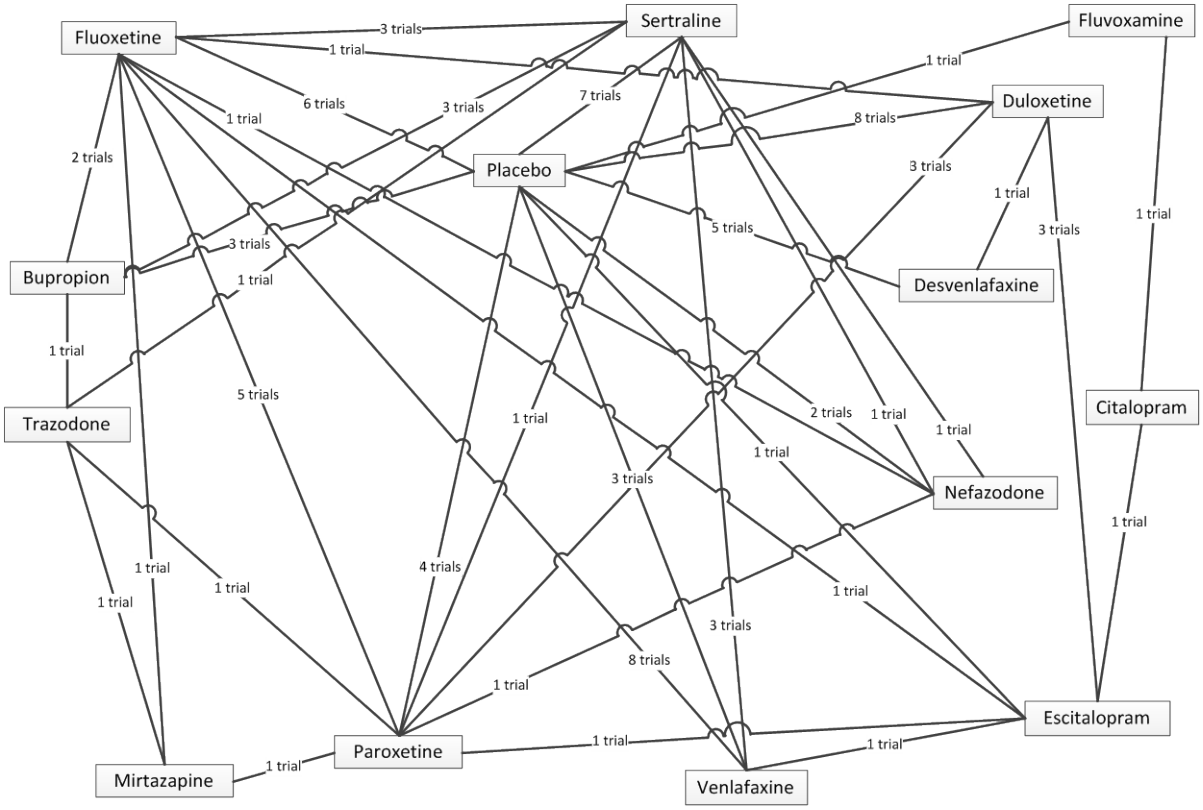
[c]Mirtazapine's mechanism of action is not clearly an SNRI, but it was grouped in this class owing to similarities.

[d]Only generic nefazodone is available in the United States.

Abbreviations: caps = capsules; SNRI = serotonin and norepinephrine reuptake inhibitor; SSRI = selective serotonin reuptake inhibitor; tabs = tablets

In our recent CER, we conducted MTC meta-analysis to derive estimates of the comparative efficacy among all SGAs for the treatment of major depressive disorder. Our primary efficacy outcome was the rate of response on the Hamilton Depression rating scale (HAM-D), defined as a 50 percent or greater improvement of scores from baseline. Figure 2 shows the evidence network that contributed data to the analysis.

Figure 2. Evidence network for mixed treatment comparison meta-analysis of second-generation antidepressants for achieving response for adults with depression



Our MTC meta-analysis found that SGAs had similar efficacy. There were some differences (based on the 95% credible intervals) for some pairwise comparisons that are likely not clinically relevant.

Biologic Disease Modifying Antirheumatic Drugs for Rheumatoid Arthritis

Our report on treatments for RA compared the benefits and harms of corticosteroids, oral and biologic disease modifying antirheumatic drugs (DMARDs) for adults with RA.[17] Nine biologic DMARDs were included in the CER (Table 2).

Table 2. Biologic DMARD treatments for rheumatoid arthritis

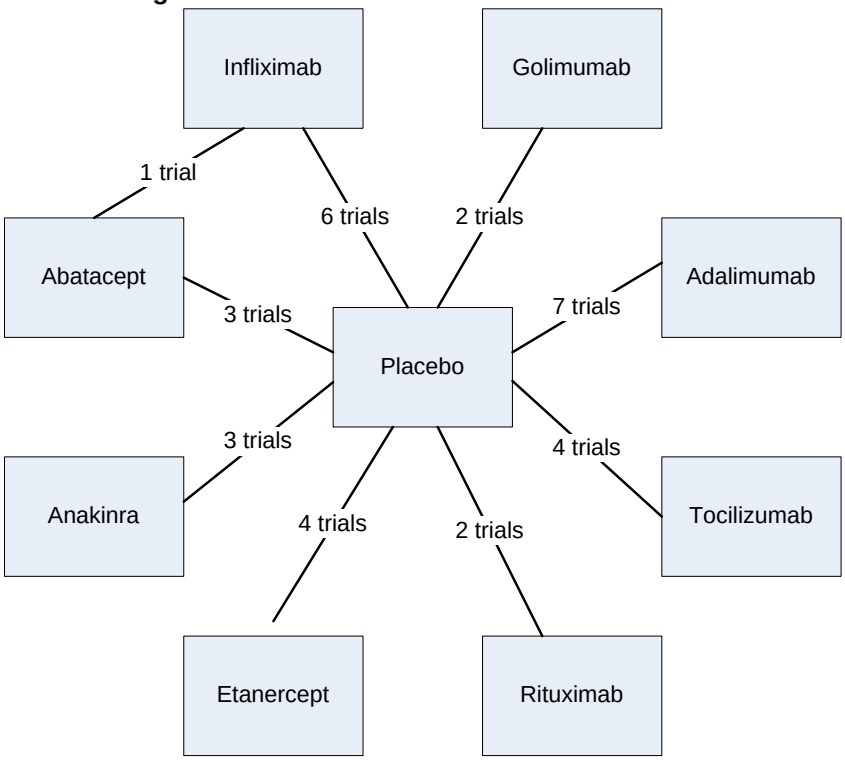
Generic Name	Manufacturer U.S. Trade Name(s)*	Injectable Supply	Usual Adult Dose
Abatacept	Bristol Myers Squibb Orencia®	250 mg powder in single-use vial, 125 mg/ml solution in a prefilled syringe	IV—Dosed according to body weight (< 60 kg = 500 mg; 60-100 kg = 750 mg; > 100 kg = 1,000 mg); dose repeated at 2 weeks and 4 weeks after initial dose, and every 4 weeks thereafter SQ—After a single IV infusion as a loading dose (as per body weight categories above), 125 mg should be given within a day, followed by 125 mg once a week
Adalimumab	Abbott Humira®	40 mg/0.8 ml pre-filled pen or syringe, 20 mg/0.4 ml prefilled syringe	SQ—40 mg every other week; may increase to 40 mg every week in patients not taking concomitant MTX
Anakinra	Amgen Kineret®	100 mg/0.67 ml syringe	SQ—100 mg/day; dose should be decreased to 100 mg every other day in renal insufficiency or end-stage renal disease
Certolizumab Pegol	UCB Cimzia®	200 mg powder for reconstitution, 200 mg/ml solution in a pre-filled syringe	SQ—Initial dose of 400 mg, repeat dose 2 and 4 weeks after initial dose, followed by 200mg every other week; for maintenance dosing, consider 40omg every 4 weeks
Etanercept	Amgen Pfizer Immunex Enbrel®	50 mg/ml autoinjector or prefilled syringe, or as two 25 mg/0.5 mL single-use prefilled syringes or free-hand vials	SQ— 50 mg once weekly with or without MTX
Golimumab	Centocor Ortho Biotech Janssen Biotech, Inc. Simponi®	50 mg/0.5 ml pre-filled syringe or autoinjector	SQ—50 mg once a month, in combination with methotrexate
Infliximab	Centocor Ortho Biotech Remicade®	100 mg lyophilized in a 20 ml vial	IV—3 mg/kg in combination with MTX at 0, 2, and 6 weeks followed by maintenance every 8 weeks thereafter; may increase to maximum of 10 mg/kg or treat as often as every 4 weeks
Rituximab	Biogen Idec / Genentech Rituxan®	100 mg/10 ml and 500 mg/50 ml vial	IV— In combination with MTX, two 1,000 mg IV infusions separated by 2 weeks (one course) every 24 weeks or based on clinical evaluation, but not sooner than every 16 weeks
Tocilizumab	Genentech / Roche Actemra®, RoActemra®	80 mg/4 ml, 200 mg/10 ml, 400 mg/20 ml vial	IV—4 mg/kg followed by an increase to 8 mg/kg based on clinical response; given every 4 weeks with or without MTX

*Listed trade names are limited to commonly prescribed U.S. products when multiple trade names are available.

Abbreviations: IV = intravenous; kg = kilogram; mg = milligram; ml, milliliter; MTX = methotrexate; SQ = subcutaneous

To compare the effectiveness of biologic DMARDs with each other, we conducted MTC meta-analyses of trials enrolling methotrexate-resistant patients with active RA. The primary efficacy outcome of our MTC meta-analysis was the American College of Rheumatology 50 percent response (ACR 50). We also conducted analyses using ACR 20 and ACR 70. Figure 3 shows the evidence network that contributed data to the analysis. The majority of evidence was from placebo-controlled trials; only one trial included a direct head-to-head comparison. Due to heterogeneity in study design of the included studies for certolizumab, it was excluded from the MTC meta-analysis. More information can be found in the full report.[17]

Figure 3. Evidence network for mixed treatment comparison meta-analysis of biologic DMARDs for achieving ACR 50 for adults with rheumatoid arthritis



Note: The total number of trials does not appear to equal 30 (the total number of studies included in the analysis) because some trials have multiple arms that were included.

Our MTC meta-analyses found higher odds of achieving ACR 50 response for etanercept compared with most other biologic DMARDs (abatacept, adalimumab, anakinra, infliximab, rituximab, tocilizumab) for methotrexate-resistant patients with active rheumatoid arthritis. The ACR 50 odds ratio range for etanercept compared with most other biologic DMARDs was 2.4 to 5.2. The differences were potentially important (based on 95% credible intervals) for etanercept compared with abatacept, adalimumab, anakinra, infliximab, rituximab, or tocilizumab, but were not when compared with golimumab. Anakinra had the lowest mean response and point estimates favored other biologic DMARDs over anakinra, but the differences were only potentially important (based on 95% credible intervals) when compared with adalimumab and etanercept for ACR 50.

Scope and Key Questions

The main objectives of this report are to contribute to the body of literature on MTC meta-analysis by examining (1a) how results of Bayesian MTC methods compare with several frequentist indirect methods for various types of outcome measures, (1b) how Bayesian MTC methods perform for different types of evidence network patterns, (2) how study level covariates can be incorporated with Bayesian MTC meta-analysis to explore heterogeneity through meta-regression, and (3) how findings of Bayesian MTC meta-analysis compare for different numbers of studies and different network pattern assumptions. We address the Key Questions (KQs) listed below.

KQ 1: How do the results of Bayesian MTC meta-analysis methods compare with those of several frequentist indirect methods? Related questions of interest included: For each of the common evidence network patterns, how do the Bayesian MTC methods compare with frequentist indirect methods? How do Bayesian MTC methods perform (e.g., precision, convergence, etc.) for different types of evidence network patterns?

KQ 2: How can meta-regression be used with Bayesian MTC meta-analysis to explore sources of heterogeneity?

KQ 3: How do findings of Bayesian MTC meta-analysis compare for different numbers of studies and network pattern assumptions?

For KQ 1, our choice of frequentist analytic methods to compare with the Bayesian MTC approach was based on our judgment regarding the methods most commonly considered by analysts conducting CERs. In addition, we selected frequentist methods with some evidence to support their validity. These included frequentist meta-regression, the Bucher method (adjusted indirect comparisons), and frequentist logistic regression. Of note, the frequentist methods used are not the analogue of the Bayesian methods implemented. We did not compare findings with the frequentist network meta-analysis method (i.e., Lumley method).[20] Our experience indicates that it is much more rarely used than the other methods, and comparisons between Bayesian MTC methods and frequentist network meta-analysis was not our intention.

We applied Bayesian MTC methods to a variety of different evidence networks using data from recent systematic reviews (of SGAs and biologic DMARDs) and using simulated data. For the first two KQs, we use the real-world bodies of literature described above (for SGAs and biologic DMARDs). For KQ 3, we use simulated data sets.

For KQ 2, we focused on clinically important issues to guide the meta-regressions. For the SGAs literature, some have questioned whether the medications are equally or less effective in older adults.[21-24] To address this question, we conducted meta-regression by assessing whether efficacy differed in trials that enrolled older adults compared with trials that enrolled adults of any age. For biologic DMARDs, some have questioned whether treatment efficacy varies by disease duration of RA.[25-27] To address this question, we conducted meta-regression using mean disease duration of subjects enrolled in each study as a continuous covariate.

Methods

In this chapter, we first describe the data from two real-world bodies of literature that we used for our analyses for Key Questions (KQs) 1 and 2. We then describe the methods used for each KQ. The focus of this report is on mixed treatment comparison (MTC) meta-analysis implemented in a Bayesian framework; for comparison, we performed adjusted indirect comparisons using four different approaches with a frequentist framework (each described below).

The National Institute for Health and Clinical Excellence (NICE) Decision Support Unit has released several Technical Support Documents (TSDs) detailing use of MTC in a Bayesian framework, including several illustrative examples with annotated WinBUGS code. We implemented the methods illustrated in TSD 2[6] and TSD 3[5] under different scenarios for the MTC meta-analyses described below. For both of our recent comparative effectiveness reviews (CERs) including MTC meta-analyses, we used WinBUGS Version 1.4.3, a Bayesian software package that uses Markov chain Monte Carlo (MCMC) techniques. WinBUGS code used for analyses is available in Appendix A.

Data Included in this Report

Dataset 1: Second-Generation Antidepressants

For this report, we extracted one binary and one continuous outcome from the evidence base of our CER on second-generation antidepressants (SGAs). Our binary outcome was treatment response as measured by at least a 50 percent improvement from baseline on the Hamilton Rating Scale for Depression (HAM-D). We recalculated response rates for each study using the number of all randomized patients as the denominator to reflect a true ITT analysis. With this approach we attempted to correct variations in results of modified ITT analyses encountered in individual studies. There were a total of 64 studies with adequate reporting of treatment response that were included in subsequent analyses. Eight additional studies were identified in the older adult populations (age 55 or older) and used in the meta-regression for KQ 2. Characteristics of the included studies and data used in the meta-analyses are listed in Appendix Table B-1. Additional description of study populations and other eligibility criteria can be found in the full report.[16]

Our continuous outcome was mean change from baseline to endpoint on the HAM-D. For studies not reporting a variance for mean change from baseline, we calculated one using the baseline and endpoint variances and assumed a correlation of 0.5. We chose a value of 0.5 for the correlation coefficient as a reasonable assumption for the similarity of baseline and endpoint values across patients, since it assumes neither a weak or strong correlation. In lieu of more complete reporting, it has been suggested that a value of 0.5 is a reasonable assumption.[28, 29] We included a total of 40 studies in the analyses for this outcome; data and characteristics of the included studies are listed in Appendix Table B-2.

Dataset 2: Biologic Disease Modifying Antirheumatic Drugs for Treatment of Rheumatoid Arthritis

We extracted one binary and one continuous outcome from the studies included in our CER on biologic DMARDs and other treatments for RA. The binary outcome considered was

treatment response as measured by achievement of ACR50 after 12 weeks of treatment. Again, we recalculated response rates for each study using the number of all randomized patients as the denominator to reflect a true ITT analysis. With this approach we attempted to correct variations in results of modified ITT analyses encountered in individual studies. A total of 31 studies covering eight biologic DMARDs with adequate reporting of treatment response were included. Characteristics and data of the included studies are shown in Appendix Table B-3 below.

We also extracted a continuous outcome, mean change from baseline in HAQ-DI (Health Assessment Questionnaire Disability Index); however, because few eligible studies reported adequate data, we were not able to perform any meaningful MTC meta-analysis on the continuous outcome.

Key Question 1: Comparison of Bayesian MTC Meta-Analysis with Frequentist Indirect Methods

Our objectives for this KQ were to examine how results of Bayesian MTC methods compare with commonly used frequentist indirect methods and how Bayesian MTC methods perform for different types of evidence network patterns. We chose four analytic methods—Bayesian MTC meta-analysis, frequentist meta-regression, the Bucher method (adjusted indirect comparisons), and frequentist logistic regression. These methods were chosen because they are among the most common approaches considered when conducting CERs. For both of our datasets (SGAs and biologic DMARDs), we compared the findings from these four methods—first, for the full networks; second, for sub-components of the full networks representing specific evidence network patterns (star, loop, one closed loop, and ladder). To compare the four analytic methods, we used several measures: (1) the proportion of drug-drug comparisons (out of the total possible number of comparisons) for which each method was unable to calculate a result, either because of model convergence issues or the lack of a common comparator; (2) the percent agreement, with findings considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for Bayesian methods; based on 95% credible intervals) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment; (3) the precision of findings—assessed by comparing the width of credible intervals and confidence intervals; and (4) kappa statistics. The kappa statistic is a measure of inter-rater agreement[30] that attempts to take into account agreement beyond chance. It can range between -1 and 1, but usually ranges from zero to 1, since kappa is negative when the observed agreement is less than chance. When the observed agreement exceeds chance agreement, kappa is positive, with its magnitude reflecting the strength of agreement. Landis and Koch propose the following as standards for strength of agreement for the kappa coefficient: <0.20=poor, 0.21-0.40=fair, 0.41-0.60=moderate, 0.61–0.80=good, and 0.81-1.00=very good.[31] For each comparison (e.g., Bayesian MTC vs. Frequentist Meta-regression) we calculated kappa statistics using SAS version 9.2. We did not calculate kappa statistics for comparisons in which either method had less than 2 levels (i.e., when all results for that method found no statistically significant or important difference for each drug-drug comparison).

Finally, we describe measures of model fit for the Bayesian MTC analyses for the full networks and for sub-components. For all dichotomous data we used odds ratios, and for all continuous data we used weighted mean differences as outcome measures.

Bayesian Mixed Treatment Comparison Meta-analysis

For all Bayesian MTC meta-analyses included in this report, we used the methods developed in TSDs 2 and 3, which detail use of the generalized linear modeling (GLM) framework for Bayesian MTC and use of meta-regression to explore sources of heterogeneity, respectively. We used random effects models in this report because in meta-analyses of randomized controlled trials, we assume the study effect is sampled from a distribution of effect sizes.[32] Because studies will not include exactly the same mix of participants or carry out the interventions in an identical way, there may be different underlying effect sizes for different studies.

GLM theory[33] allows for likelihood based statistical inference. It is also a flexible approach because it can be constructed to model data arising from a large range of distributions within the exponential family, allowing us to model various binary and continuous outcomes that we would encounter in conducting CERs. GLM theory can be implemented in both frequentist and Bayesian frameworks. Speigelhalter et al., 2004[34] make the case for a Bayesian framework, stating that such an approach is more flexible, efficient, and useful. MTC could in theory be implemented in a frequentist framework, but the literature reflects a preference for a Bayesian framework, in part due to the availability of WinBUGS code that has been developed for this purpose.[6, 35] Reasons to use a Bayesian framework (rather than a frequentist approach) include inferential superiority and modeling flexibility. Whether Bayesian approaches have inferential superiority is a contentious subject and may have little to do with the reason the Bayesian MTC methods are used more widely. Modeling flexibility, however, is likely a major reason the Bayesian approach is more popular. With WinBUGS, one has simply to specify the likelihood and the prior distributions; the estimation is handled by MCMC methods which are now fairly robust. With maximum likelihood estimation using a frequentist approach, some of the likelihoods can be very challenging to maximize and perhaps amenable to approximate solutions only.

In a Bayesian framework, specification of prior distributions accompanies the specification of the likelihoods, in addition to the data. The complete model specification is detailed below. For this report, binary outcomes are treatment response measured by the number of people achieving a pre-specified level in each arm. We used flat (noninformative) prior distributions throughout the report, which allow the data to drive the posterior distributions, in absence of more informative priors. This choice follows the recommendations laid out in TSD 2[6] and in absence of rationale for specifying informative priors with these data. For all analyses, we modeled study effect and treatment effect parameters by noninformative (flat) prior distributions that were normal (0, 10,000). For the heterogeneity of the random-effects model, we used a uniform prior distribution centered at zero with sufficiently large variance. Unless otherwise specified, we discarded the first 20,000 simulations to allow for model convergence, and we used an additional 100,000 simulations in estimating the posterior probabilities. Satisfactory convergence was verified by trace plots and monitoring the Monte Carlo error. If multi-arm studies were available, we included the appropriate adjustments to the likelihood to account for correlations between the treatment differences.

Bayesian Mixed Treatment Comparison: Model Specification

Most of the data used in this report consisted of outcomes that measured the number of events, in all cases treatment response, occurring out of the total number of patients. These types of data were assumed to arise from a Binomial likelihood. The likelihood was specified as

$$r_{i,k} \sim Binomial(p_{i,k}, n_{i,k})$$

where r_{ik} is the number achieving treatment response out of the total number of patients in each arm and p_{ik} is the probability of treatment response, and where i denotes trial number and k denotes each the arm of the trial. For this likelihood, the logit link function was used to map the probability of treatment response onto plus and minus infinity. The logit link was specified as

$$\text{logit}(p_{i,k}) = \mu_i + \delta_{i,k} I_{(k \neq 1)}$$

The trial-specific log-odds ratios from a random effect model come from a common distribution, $\delta_{ik} \sim \text{N}(d_k, \sigma^2)$, where d represents the relative treatment effect and σ is the common variance term.

In all cases, parameters were given vague or noninformative prior distributions. The choice of the uniform prior between 0 and 5 reflects a between-study variance that allows for a wide range of treatment effects.

Prior specification:

$$\mu_i \sim \text{Normal } (0, 10000)$$
$$d_k \sim \text{Normal } (0, 10000)$$
$$\sigma \sim \text{Uniform } (0, 5)$$

For continuous outcomes, the data were assumed to arise from a normal likelihood, and under generalized linear modeling theory, the model specified above becomes a normal likelihood with the identity link function. The observed data $y_{i,k}$ arises from a normal distribution with mean $\theta_{i,k}$ and variance $se_{i,k}^2$.

$$y_{i,k} = N(\theta_{i,k}, se_{i,k}^2)$$

For these data, the prior specifications were identical to the ones used with the dichotomous data.

For KQ2, in order to introduce a dichotomous study-level covariates to the logistic regression model above, the model was simply extended to include a covariate x_i. The model specification and resulting WinBUGS code were taken from the Decision Support Unit's Techincal Support Document 3. Again, the subscript i denotes trials and the subscript k refers to the arms within a trial. The full model structure is specified below.

$$r_{i,k} \sim Binomial(p_{i,k}, n_{i,k})$$
$$\text{logit}(p_{i,k}) = \mu_i + \delta_{i,k} + (\beta_k - \beta_1) x_i$$
$$\beta_1 = 0; \ \delta_{i,1} = 0$$

Prior specification:

In addition to the above, now a vague normal prior on the common covariate effect

$$\beta_k \sim \text{Normal } (0, 10000)$$

Again, the extension to introducing a continuous covariate results in the same model above, but to improve the efficiency of estimation, the continuous covariate was centered around its mean, as shown below. All other pieces of the model specification remained the same. Changes to the logistic model are shown below.

$$\text{logit}(p_{i,k}) = \mu_i + \delta_{i,k} + (\beta_k - \beta_1)(x_i - mean(x_i))$$

Model Statistics

When conducting the Bayesian MTC meta-analyses, we output several statistics to compare relative efficacy and assess model fit, as described below.

Outcome Measures

We assessed relative efficacy between treatments with odds ratios and 95 percent credible intervals for dichotomous data, and with mean differences and 95 percent credible intervals for continuous data. We also calculated the probability that each treatment was the best, by ranking the drugs on a relative scale. Because we used random effects models, we also output the estimate of the between-studies standard deviation in each scenario.

Model Fit

We assessed model fit with the Deviance Information Criterion (DIC) and the posterior mean of the total residual deviance.[36] Deviance measures the fit of the model to the data using the likelihood function. A good model fit is indicated by a total residual deviance approximately equal to the number of data points available. The DIC is a statistic that measures Bayesian model fit and penalizes the deviance by the model complexity. When comparing two DIC values, a difference of 5 or more is regarded as a meaningful difference.[37]

Frequentist Indirect Comparisons

We compared the results from Bayesian MTC meta-analyses with three frequentist indirect methods for binary data and two frequentist indirect methods for continuous data. We used only two for continuous data because logistic regression cannot be used with continuous data. We chose these methods because they are among the most common indirect analyses currently used in CERs and because they have been validated with data simulation.[3]

Frequentist Meta-Regression

For the frequentist meta-regression approach, we followed a two-step process to indirectly estimate the comparative efficacy of two drugs. We first conducted random effects meta-analyses as proposed by DerSimonian and Laird,[4] combining two drugs of interest to estimate an overall pooled effect. We assumed that i of a total of n studies provided an estimate y_i of the effect (log odds ratio for dichotomous outcomes and weighted mean differences for continuous outcomes):

$y_i = \theta_i + \upsilon_i + \varepsilon_i$.
$\upsilon_i \sim N(0, \tau^2)$
$\varepsilon_i \sim N(0, \sigma_i^2)$

where τ^2 is the between study variance, σ_i the standard error of the estimate y_i. The model assumes that υ_i and ε_i are uncorrelated. We implemented this method using the Stata "metan" command with the "random" option. The between-study variance τ^2 is estimated by the methods of moments: $\frac{\tau^2 = Q - df}{C}$

where

$$\tau^2 = \frac{Q - df}{C}$$

df=k-1

where k is the number of studies, and

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

Second, we estimated the comparative treatment effects between two drugs with random effects meta-regression. The unit of analysis for meta-regression was a study. The predictor in the regression model was a binary variable indicating the presence or absence of each of two drugs of interest. Random effects meta-regression extends random effects meta-analysis by replacing the mean θ_i with a linear predictor $x_i\beta$. Random effects meta-analysis allows for residual heterogeneity by assuming that the true effects follow a normal distribution around the linear predictor:[38]

$y_i = x_i\beta + \upsilon_i + \varepsilon_i$.
$\upsilon_i \sim N(0, \tau^2)$
$\varepsilon_i \sim N(0, \sigma_i^2)$

We implemented this method using the Stata "metareg" command. Stata version 11 was used for all frequentist meta-regression analyses.

Bucher Method

For another adjusted indirect comparisons approach, we used methods proposed by Bucher and colleagues.[2] To derive indirect comparisons of two treatments, the method compares the magnitude of treatment effects of two interventions relative to a common comparator. For binary data, we calculated the pooled odds ratios for each drug of interest relative to a common comparator using random effects meta-analyses (as proposed by DerSimonian and Laird). For continuous data we used weighted mean differences.

As outlined above in the section on meta-regression, for the meta-analyses we assumed that i of a total of n studies provided an estimate y_i of the effect (log odds ratio for dichotomous outcomes and weighted mean differences for continuous outcomes):

$y_i = \theta_i + \upsilon_i + \varepsilon_i$.
$\upsilon_i \sim N(0, \tau^2)$
$\varepsilon_i \sim N(0, \sigma_i^2)$

where τ^2 is the between study variance, σ_i the standard error of the estimate y_i. The model assumes that υ_i and ε_i are uncorrelated.

The between-study variance τ^2 is estimated by the methods of moments, as outlined under meta-regression.

For dichotomous outcomes, the difference between two log odds ratios (with a common comparator) provides an adjusted estimate for the indirect comparison of agents:

$$\ln(OR_{ind}) = \ln(OR_{ac}) - \ln(OR_{bc})$$

Likewise, the sum of the variance of the two meta-analyses is the basis of the confidence interval for this calculation.

$$\mathrm{Var}(lnOR_{ind}) = Var(\ln OR_{ac}) + Var(\ln OR_{bc})$$

For continuous outcomes we used the same analytic approach without conversion to a logarithmic scale. The effect measure of choice for continuous outcomes was the weighted mean difference of the treatments relative to a common comparator. We conducted all analyses with Comprehensive MetaAnalysis version 2.2.050 (Biostat, Englewood NJ).

Logistic Regression Modeling

For the indirect comparisons using logistic regression, we incorporated the approach outlined by Glenny et al. (2005).[3] We constructed datasets with the number of participants responding to the treatment and the number of participants not responding to the treatment matching those reported in each study. In other words, if a study reported 100 of 300 patients responded to treatment A and 150 of 300 patients responded to treatment B, we constructed a dataset with a total of 600 patients (100 responders and 200 non-responders receiving treatment A and 150 responders and 150 non-responders receiving treatment B). We then estimated logistic regression models using PROC GLIMMIX in SAS version 9.2 with a dichotomous outcome of response vs. no response to treatment and including treatment as a fixed effect and study as a random effect.

PROC GLIMMIX is used for computing generalized linear mixed effect models and can be applied to various types of outcome variables. The general form of the model is:

$$E(Y|\gamma) = g^{-1}(X\beta + Z\gamma)$$

where Y = outcome variable,
X = fixed effects,
β = fixed effect parameters,
γ = random effects,
Z = random effect parameters, and
g = link function.

In this model, E(Y| γ) represents the expected value for the outcome based on the model. In our analyses, the outcome was response to treatment (yes vs. no) with treatment as fixed effect and study as a random effect as shown below:

$$E(Response\ to\ treatment|\gamma) = g^{-1}(\beta_0 + \beta_1 Treatment_1 \ldots + \beta_i Treatment_i + \gamma_1 Study_1 \ldots . + \gamma_j Study_j)$$

where i is the number of treatments and j is the number of studies.

Because the outcome is dichotomous, we selected a logit link function and when applied to the equation, the final model is:

$$E(Response\ to\ treatment|\gamma) = \frac{1}{1 + e^{-(\beta_0 + \beta_0 + \beta_1 Treatment_1 \ldots + \beta_i Treatment_i + \gamma_1 Study_1 \ldots . + \gamma_j Study_j)}}$$

A binomial variance function was specified along with an unstructured covariance matrix which permits the estimation of all elements of the variance/covariance matrix without constraints. The indirect comparisons were estimated within these models by applying contrast statements to compute odds ratios comparing response between two treatments of interest.

Handling of Multi-Arm Studies in Frequentist Analyses

None of the frequentist indirect methods used in this report accounted for correlations between treatment responses within multi-arm studies. This is an advantage of the Bayesian MTC methods when multi-arm studies are included, as the correction to the likelihood is easily implemented in the available code. However, in most cases this was not an issue because multi-arm studies were generally not included when conducting the indirect frequentist analyses.

Choice of Network Patterns

As part of KQs 1 and 2, we aimed to investigate how Bayesian MTC methods perform for different network patterns. An evidence network's geometry may be shaped by clinical context, regulatory pressure, or other factors. The evidence network may change over time, as new drugs are added to the market or different comparators are chosen.[18] We chose four simple network patterns as the basis for comparisons within this report: star, closed loop, loop plus one, and ladder (see Introduction). These four patterns reflect different scenarios that exist in the context of real-world data. Within each of the real-world datasets used, we selected a subset of the studies for each of the four patterns. Although there were many different subsets of data available for each pattern that we could have selected, we selected examples that maximized the amount of data available for each sub-network. We selected one of each type of network pattern for each of the datasets. Due to limited time and resources, we could not evaluate every possible network pattern within each dataset. Appendix C illustrates the network patterns evaluated. Other choices could have been made, such as one reflecting the entry of drugs into the market or chronologically by publication date, but for the purposes of this report, we believe that an approach maximizing the amount of data available is the most useful.

Key Question 2: Meta-Regression

Our objective for this KQ was to introduce study level covariates in the models used with Bayesian MTC meta-analysis to explore heterogeneity, attempting to answer two clinically

important questions—one for each of our datasets. For the SGAs dataset, we conducted meta-regression by assessing whether efficacy differed in trials that enrolled older adults compared with trials that enrolled adults of any age. For the biologic DMARDs dataset, we conducted meta-regression using mean disease duration of subjects enrolled in each study as a continuous covariate to determine whether treatment efficacy varies by disease duration of RA. The two analyses involved exploring two different types of covariates in the meta-regression: one relates to exploring subgroup effects with a binary covariate (for SGAs), and the other to exploring interaction effects with a continuous covariate (disease duration of RA).

There is often a need to explore heterogeneity in treatment effects in terms of another variable (e.g., an effect modifier), and many software packages facilitate meta-regression for pair-wise meta-analysis. But, until recently, there has not been a readily available process for Bayesian MTC in WinBUGS. The NICE Decision Support Unit published TSD 3[5] to illustrate the use of different types of meta-regression within Bayesian MTC. Bayesian MTC allows us to incorporate all the available evidence into one analysis, which, combined with the ability to explore potential effect modifiers, can be a powerful tool when conducting CERs.

We used the methods developed in TSD 3, which build on the generalized linear modeling framework established in TSD 2. For exploring subgroup effects through a meta-regression, a trial-level binary indicator was created to indicate study population. This indicator variable was added to the random effects logistic regression model used in KQ1. In this way, the interaction term describes the effect of the population on the outcome. In the second example, a continuous covariate was added to the same random effects logistic regression model, assuming a common covariate effect for each treatment. The continuous covariate was centered at its mean value to improve model convergence. The magnitude and direction of the interaction term was examined in each example, along with its effect on the odds of treatment response in each case.

Meta-Regression With a Subgroup Indicator Covariate: Efficacy of SGAs and Older Adults

From our report on SGAs, eight trials provided data on efficacy in older adult patients. We hypothesized that differences in efficacy may exist between studies enrolling older adults (≥55 years) and those enrolling adults of any age. We explored this by introducing a covariate to the model that indicated the subgroup population. The analysis included 72 trials; 64 trials addressing efficacy in the adult population and eight trials conducted in the older adult population. The outcome was treatment response as measured by 50 percent or greater improvement from baseline on the HAM-D. We intended to test the interaction effect and compare the odds ratios for the treatments within each population. One advantage of this type of meta-regression is that it allows for the exploration of differences within one model, rather than performing multiple analyses separately and comparing results qualitatively.

Meta-Regression With a Continuous Covariate: Efficacy of Biologic DMARDs and Disease Duration of Rheumatoid Arthritis

Our recent CER on biologic DMARDs included a total of 31 studies in the MTC meta-analysis (see descriptions above and data of the included studies provided in Appendix Table B-3). The outcome was response as measured by ACR50 after a minimum treatment period of 12 weeks. The trials included RA patients who failed methotrexate, a first-line therapy, but were naïve to treatment with a biologic DMARD. One trial-level factor hypothesized to have an effect

on treatment response is mean disease duration of rheumatoid arthritis. Longer disease durations are hypothesized to be associated with higher levels of treatment response. We aimed to test the effects of this covariate (mean disease duration) on our findings (with etanercept yielding greater treatment response than most other biologic DMARDs: see Introduction).

Key Question 3: Methods Section for Simulation Study

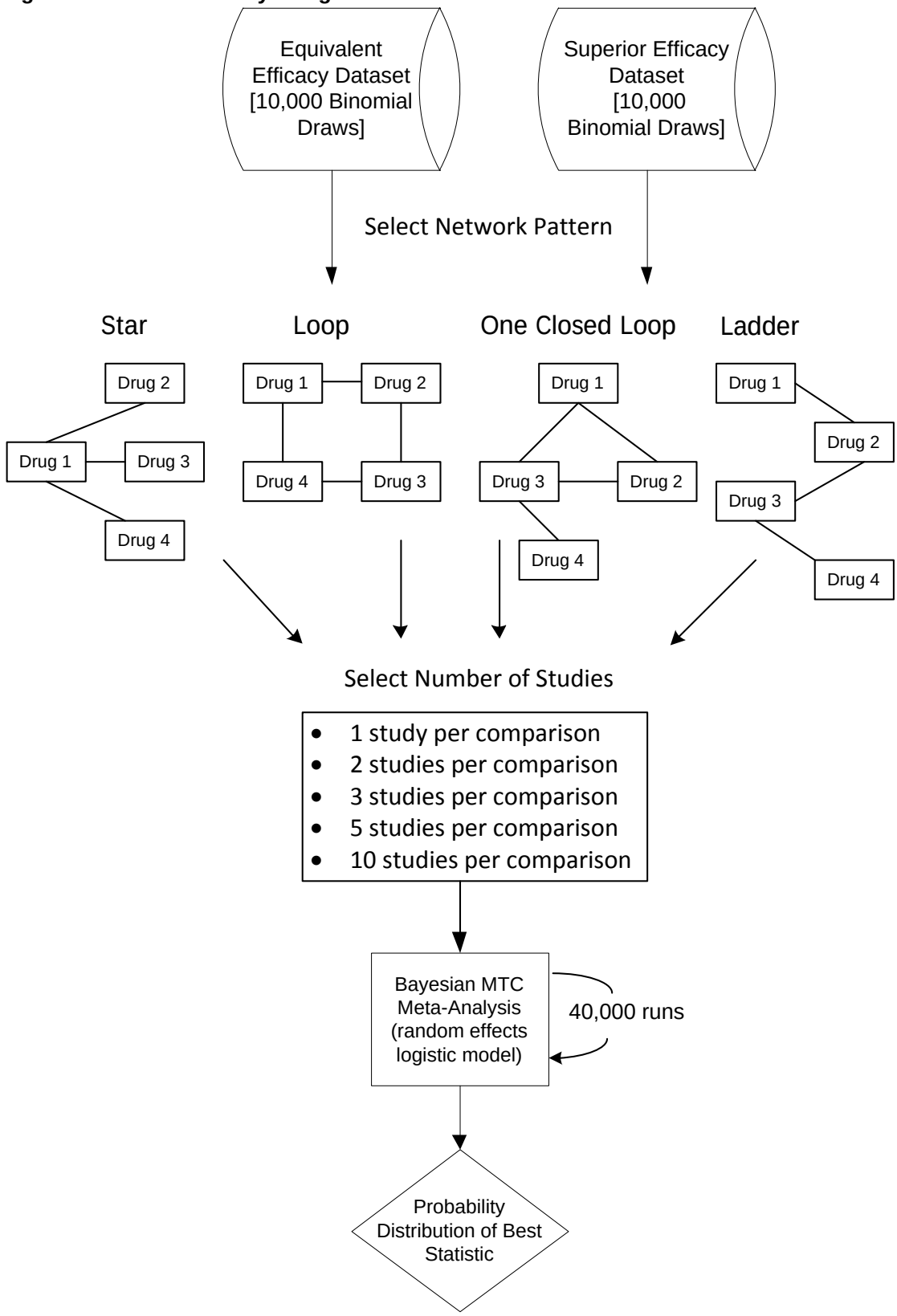
Because of inherent limitations when using real-world data, we proposed a simulation study to investigate the role of the number of studies available for each comparison on the MTC meta-analysis model's ability to detect true relative efficacy under different network patterns. The motivation for a simulation study arose from uncertainty about how the number of studies available for each comparison affects the ability of the model to produce valid results, and whether this varied under different network patterns. In performing Bayesian MTC meta-analyses for CERs, the question of whether there are enough studies to make valid conclusions about comparative efficacy often arises. Many times, only one study is identified that links two treatments (or a treatment and placebo) together in a network. We hypothesized that findings may vary depending on the underlying data structure (i.e., network pattern).

In order to assess the ability of the Bayesian MTC method to produce valid results under the four network patterns, we first created simulated datasets comprising two scenarios for comparative efficacy. Simulated datasets will lend flexibility in choosing the number of studies available for each comparison, and also allow us to determine the true comparative efficacy a priori. The basic process of the simulation is illustrated in Figure 4, below. We restricted the simulations to four treatments in each case, and set the sample size *within* each study to be 100 patients per treatment arm. Our outcome was a dichotomous measure of treatment response (1=response, 0=no response).

Two master simulated datasets were created to test the method in one case where analyses should not find important differences and in another case where analyses should find important differences. The first dataset is a scenario of equivalent efficacy of three drugs compared with placebo. A dataset of 10,000 'studies' and four treatments was created by sampling from the binomial distribution so that placebo had a mean response of 0.10, and drugs 2 through 4 had a mean response of 0.50. In this way, drugs 2 through 4 have equivalent efficacy, but all have greater efficacy than placebo.

The second of the two datasets is a scenario of superior efficacy of one drug compared with three other drugs. Again, a dataset of 10,000 'studies' and four treatments was created by sampling from the binomial distribution so that drugs 1 through 3 had a mean response of 0.20, and drug 4 had a mean response of 0.80. In this case, drug 4 is superior to the other drugs in the network, which all have equal efficacy compared with each other.

Figure 4. Simulation study design



Next, to address variations in the number of studies available for each comparison, we selected five cases: one, two, three, five, and 10 studies per comparison. That is, the link connecting any two drugs is based on one, two, three, etc., studies. We fixed this number so that the overall number of studies per link is the same for each drug comparison. We sampled successive studies from the two master datasets, without replacement, to produce 1,000 sample datasets for each of the four network patterns (star, loop, loop plus one, and ladder) and the five cases, resulting in a total of 40,000 simulated datasets. The datasets were then submitted to the Bayesian MTC meta-analysis model used in KQ1. In each model run, there was a burn-in of 5,000 iterations, followed by 15,000 iterations from which we monitored and output the probability that each drug was the best (i.e., most efficacious). The output of 1,000 'best' statistics for each pattern by case (determined by number of studies for each comparison) formed a distribution to allow us to assess the model estimates versus the pre-determined treatment efficacy. To estimate bias, we computed the difference between the percentage of times a drug was considered most effective based on the models and the expected percentage based on the assumptions used to generate the data. The SAS and WinBUGs code used to generate the master datasets and Bayesian models can be found in Appendix A.

Results

Key Question 1. Bayesian Mixed Treatment Comparison (MTC) Methods Compared With Frequentist Indirect Methods, and Performance for Different Types of Evidence Network Patterns

We organized this section in two main parts: (1) comparison of results of Bayesian MTC methods with those of various frequentist methods for the full evidence networks and (2) comparison of and performance for different types of evidence network patterns. For the second part, we divided the results into two sub-sections—one comparing results of Bayesian MTC methods with those of various frequentist methods for sub-components (i.e., evidence network patterns) of the full networks and one presenting results of Bayesian MTC methods for the full networks and those of Bayesian MTC methods for sub-components of the full networks. Within each part, we first address results using data from our second-generation antidepressants (SGAs) report and then results using data from our report on treatments for rheumatoid arthritis (RA).

Comparison of Bayesian MTC Results with those of Various Frequentist Analyses for the Full Networks

In this section, we provide results of comparisons between Bayesian MTC methods and three frequentist methods for the full networks illustrated in Appendix C. Tables in Appendix D provide detailed results for each analysis. Tables in Appendix E provide a comparison of precision of findings from the various analyses (determined by width of the 95% credible interval or confidence interval), with the darkest shading indicating the most precise result and the lightest indicating the least precise for each drug-drug comparison. The Tables in this section provide a summary and some comparison of the data in Appendix D and Appendix E.

Second-generation Antidepressants: Response (Binary Outcome)

Appendix Table D-1 provides results of our analyses for the full SGA network for each drug-drug comparison for Bayesian MTC meta-analyses, frequentist meta-regression, the Bucher method, and frequentist logistic regression (Appendix D). For 15 out of 78 drug-drug comparisons, neither frequentist meta-regression nor the Bucher method produced a result either because no studies included a common comparator or because an insufficient number of studies included a common comparator for the program to run the analysis.

Table 3 summarizes the number of comparisons for which each frequentist method was unable to produce a result (the Bayesian MTC method produced results for all 78 comparisons), the percent agreement and kappa between each frequentist method and the Bayesian MTC method, and comparative precision. The three frequentist methods were unable to produce a result for 0 percent (0/78), 32 percent (25/78), and 45 percent (35/75) of the drug-drug comparisons (for the logistic regression, meta-regression, and Bucher methods, respectively). Logistic regression always produced the most precise result and Bayesian MTC was the next most precise for all but one drug-drug comparison (Appendix Table E-1).

Table 3. Comparison of findings for SGA response between Bayesian MTC meta-analysis findings and those of frequentist methods: agreement and kappas

Outcome	Bayesian MTC	Meta-Regression	Bucher Method	Logistic Regression
No result produced[a]	0 of 78 comparisons	25 of 78 comparisons	35 of 78 comparisons	0 of 78 comparisons
% agreement with Bayesian MTC[b]	NA	94.3 (50/53)	93.0 (40/43)	100
Kappa[c]	NA	NA	NA	1.00
Precision	2nd most precise	Least precise	3rd most precise	Most precise

MTC= mixed treatment comparison; NA=not applicable

[a]No result produced indicated the number of drug-drug comparisons for which the method was unable to produce a result either due to no studies with a common comparator or an insufficient number of studies with a common comparator.

[b]Percent agreement calculated only for drug-drug comparisons that both methods were able to produce a result. For example, meta-regression did not produce a result for 25 of 78 drug-drug comparisons. Therefore the percent agreement was calculated using the 53 comparisons for which both methods produced a result. Results were considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for Bayesian MTC) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment.

[c]Kappa was calculated only for comparisons when both methods were able to produce a result, as with percent agreement. Not applicable (NA) indicates that the statistical program was unable to calculate a kappa because of insufficient data.

For the results of the three drug-drug comparisons that were not in agreement between the Bayesian MTC and the meta-regression and Bucher methods (duloxetine vs. escitalopram, escitalopram vs. fluoxetine, and fluoxetine vs. venlafaxine), the Bayesian MTC meta-analysis found a potentially important difference (based on 95% credible intervals) between treatments, whereas the other two methods did not. In addition, the results of the Bayesian MTC had greater precision than the other two methods for all three comparisons. Point estimates were very similar for the three methods for one of the comparisons (fluoxetine vs. venlafaxine, ORs ranged from 0.75 to 0.77) but not for the other two comparisons (duloxetine vs. escitalopram: ORs 0.74 vs. 1.23 vs. 1.19, respectively; escitalopram vs. fluoxetine: ORs 1.44 vs. 0.94 vs. 0.94).

Second-generation Antidepressants: Mean Change in HAM-D (Continuous Outcome)

Appendix Table D-2 provides results of our analyses for the full SGA network for each drug-drug comparison for Bayesian MTC meta-analyses, frequentist meta-regression, and the Bucher method (Appendix D). There are no results from frequentist logistic regression for this outcome because it is a continuous outcome.

For 33 out of 78 drug-drug comparisons, neither of the frequentist methods produced a result either because no studies included a common comparator or because an insufficient number of studies included a common comparator for the program to run the analysis. Table 4 summarizes the number of comparisons for which each frequentist method was unable to produce a result (the Bayesian MTC method produced results for all 78 comparisons) as well as the percent agreement between each frequentist method and the Bayesian MTC method. Bayesian MTC meta-analysis produced the most precise result for 67 percent of the comparisons; the Bucher method was the most precise for 33 percent of the comparisons (Appendix Table E-2).

Table 4. Comparison of findings for mean change in HAM-D between Bayesian MTC meta-analysis findings and those of frequentist methods: agreement and kappas

Outcome	Bayesian MTC	Meta-Regression	Bucher Method
No result produced[a]	0 of 78 comparisons	46 of 78 comparisons	33 of 78 comparisons
% agreement with Bayesian MTC[b]	NA	100 (32/32)	84.4 (38/45)
Kappa[c]	NA	NA	NA
Precision	Most precise	Least precise	2nd most precise

HAM-D=Hamilton Depression rating scale; MTC=mixed treatment comparison; NA=not applicable

[a]No result produced indicated the number of drug-drug comparisons for which the method was unable to produce a result either due to no studies with a common comparator or an insufficient number of studies with a common comparator.

[b]Percent agreement calculated only for drug-drug comparisons that both methods were able to produce a result. Results were considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for Bayesian MTC) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment.

[c]Not applicable (NA) indicates that the statistical program was unable to calculate a kappa because of insufficient data.

For the results of the seven drug-drug comparisons that were not in agreement between the Bayesian MTC and the Bucher methods, the Bucher method found a statistically significant difference between treatments, whereas the Bayesian MTC method did not find an important difference (based on the 95% credible interval). The seven comparisons were duloxetine versus mirtazapine, duloxetine versus venlafaxine, escitalopram versus trazodone, escitalopram versus venlafaxine, mirtazapine versus venlafaxine, paroxetine versus venlafaxine, and sertraline versus venlafaxine. The Bayesian MTC method usually produced a more precise result (based on comparison of credible intervals with confidence intervals), but the point estimates differed in magnitude and sometimes in the direction (three of the seven) of effect (i.e., point estimates trended in opposite directions, favoring different treatments).

Biologic Disease Modifying Antirheumatic Drugs for Rheumatoid Arthritis: ACR50 (Binary Outcome)

Appendix Table D-3 provides results of our analyses for the full biologic DMARDs network for each drug-drug comparison for Bayesian MTC meta-analyses, frequentist meta-regression, the Bucher method, and frequentist logistic regression (Appendix D). For all 28 drug-drug comparisons, all four methods were able to produce results. Table 5 summarizes the percent agreement and kappa between each frequentist method and the Bayesian MTC method. Logistic regression produced the most precise result for 89 percent of the comparisons; the Bayesian MTC meta-analysis produced the third most precise result for 86 percent of the comparisons (Appendix Table E-3).

Table 5. Comparison of findings for ACR50 between Bayesian MTC meta-analysis findings and those of frequentist methods: agreement and kappas

Outcome	Bayesian MTC	Meta-Regression	Bucher Method	Logistic Regression
No result produced	0 of 28 comparisons	0 of 28 comparisons	0 of 28 comparisons	0 of 28 comparisons
% agreement with Bayesian MTC[a]	NA	96.4 (27/28)	92.9 (26/28)	82.1 (23/28)
Kappa	NA	0.90	0.83	0.62
Precision	3rd most precise	Least precise	2nd most precise	Most precise

[a]Results were considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for Bayesian MTC) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment.

Abbreviations: ACR50=American College of Rheumatology 50 percent response; MTC=Mixed treatment comparison; NA=not applicable

For the results of the one drug-drug comparison (adalimumab vs. anakinra) that was not in agreement between the Bayesian MTC meta-analysis and meta-regression, the Bayesian MTC method found a potentially important result favoring adalimumab over anakinra (OR, 1.88, 95% CrI 1.01 to 3.98). Meta-regression found a similar point estimate but with a wider confidence interval not reaching statistical significance (OR, 1.95, 95%, CI, 0.82 to 4.61).

For the results of both drug-drug comparisons (anakinra vs. golimumab and etanercept vs. golimumab) that were not in agreement between the Bayesian MTC and the Bucher method, the Bucher method found a statistically significant difference between treatments, whereas the Bayesian MTC did not find an important difference. Point estimates for both were fairly similar and in the same direction, but the confidence intervals were more narrow (reaching statistical significance) for the Bucher method.

For the results of the five drug-drug comparisons that were not in agreement between the Bayesian MTC and logistic regression, logistic regression found a statistically significant difference between treatments, whereas the Bayesian MTC did not find an important difference. The five comparisons were abatacept versus anakinra, anakinra versus golimumab, anakinra versus infliximab, anakinra versus rituximab, and anakinra versus tocilizumab. Point estimates were fairly similar and were in the same direction for all five for the Bayesian MTC meta-analysis and logistic regression, but the confidence intervals were more narrow (reaching statistical significance) for logistic regression.

Performance for Different Types of Evidence Network Patterns

Comparison of Bayesian MTC Results with those of Various Frequentist Analyses for Sub-components of the Full Networks

In this section, we provide results of comparisons between Bayesian MTC methods and three frequentist methods for the sub-components of the full networks illustrated in Appendix C. These sub-components represent at least one of each of the following network patterns: placebo star, loop, one closed loop, and ladder. Tables in Appendix D provide detailed results for each analysis, by network pattern. Tables in Appendix E provide a comparison of precision of findings from the various analyses (determined by width of the 95% credible interval or confidence interval), with the darkest shading indicating the most precise result and the lightest

indicating the least precise for each drug-drug comparison. The Tables in this section provide a summary and some comparison of the data in Appendix D and Appendix E.

Second-generation Antidepressants: Response (Binary Outcome)

Appendix Tables D-4 through D-7 provide results of our analyses for each of the network patterns (sub-components of the full network) selected, including a placebo star, loop, one closed loop, and ladder for each drug-drug comparison for Bayesian MTC meta-analyses, frequentist meta-regression, the Bucher method, and frequentist logistic regression (Appendix D). Table 6 summarizes the number of comparisons for which each frequentist method was unable to produce a result as well as the percent agreement and kappa between each frequentist method and the Bayesian MTC method, by network pattern. On average, logistic regression produced the most precise results for the star, one closed loop, and ladder, whereas the Bayesian MTC meta-analysis produced the most precise results for the loop (Appendix Tables E-4 through E-7).

Table 6. Comparison of findings for SGA response between Bayesian MTC meta-analysis findings and those of frequentist methods for various network patterns

Network Pattern	Outcome	Bayesian MTC	Meta-Regression	Bucher Method	Logistic Regression
Star	No result produced[a]	0 of 45 comparisons	3 of 45 comparisons	2 of 45 comparisons	0 of 45 comparisons
	% agreement with Bayesian MTC[b]	NA	100	100	100
	Kappa[c]	NA	NA	NA	NA
	Precision	3rd most precise	Least precise	2nd most precise	Most precise
Loop	No result produced[a]	0 of 3	0 of 3	0 of 3	0 of 3
	% agreement with Bayesian MTC[b]	NA	66.7 (2/3)	66.7 (2/3)	66.7 (2/3)
	Kappa[c]	NA	NA	NA	NA
	Precision	Most precise	Least precise	3rd most precise	2nd most precise
One closed loop	No result produced[a]	0 of 32	25 of 32	25 of 32	0 of 32
	% agreement with Bayesian MTC[b]	NA	85.7	85.7	93.8
	Kappa[c]	NA	NA	NA	0.6364
	Precision	2nd most precise	Least precise	3rd most precise	Most precise
Ladder	No result produced[a]	0 of 55	39 of 55	36 of 55	0 of 55
	% agreement with Bayesian MTC[b]	NA	100	100	96.4
	Kappa[c]	NA	NA	NA	0.6474
	Precision	3rd most precise	Least precise	2nd most precise	Most precise

[a]No result produced indicated the number of drug-drug comparisons for which the method was unable to produce a result either due to no studies with a common comparator or an insufficient number of studies with a common comparator.

[b]Percent agreement calculated only for drug-drug comparisons that both methods were able to produce a result. Results were considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for Bayesian MTC) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment.

[c]Kappa was calculated only for comparisons when both methods were able to produce a result, as with percent agreement. Not applicable (NA) indicates that the statistical program was unable to calculate a kappa because of insufficient data (because there were less than 2 levels).

Abbreviations: MTC=mixed treatment comparison; NA=not applicable; SGA=second-generation antidepressant

Second-Generation Antidepressants: Mean Change in HAM-D (Continuous Outcome)

Appendix Tables D-8 through D-11 provide results of our analyses for each of the network patterns (sub-components of the full network) selected, including a placebo star, loop, one closed loop, and ladder for each drug-drug comparison for Bayesian MTC meta-analyses, frequentist meta-regression, and the Bucher method (Appendix D). There are no results from frequentist logistic regression for this outcome because it is a continuous outcome.

Table 7 summarizes the number of comparisons for which each frequentist method was unable to produce a result (the Bayesian MTC method produced results for all comparisons) as well as the percent agreement and kappa between each frequentist method and the Bayesian MTC method, by network pattern. On average, logistic regression produced the most precise results for the star and ladder, the Bayesian MTC meta-analysis produced the most precise results for the loop, and meta-regression produced the least precise results for all network patterns (Appendix Tables E-8 through E-11).

Table 7. Comparison of findings for mean change in HAM-D between Bayesian MTC meta-analysis findings and those of frequentist methods for various network patterns

Network Pattern	Outcome	Bayesian MTC	Meta-Regression	Bucher Method
Star	No result produced[a]	0 of 21 comparisons	1 of 21 comparisons	0 of 21 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (20/20)	81 (17/21)
	Kappa[c]	NA	NA	NA
	Precision	2nd most precise	Least precise	Most precise
Loop	No result produced[a]	0 of 3 comparisons	2 of 3 comparisons	0 of 3 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (1/1)	100 (3/3)
	Kappa[c]	NA	NA	NA
	Precision	Most precise	Least precise	2nd most precise
One closed loop	No result produced[a]	0 of 21 comparisons	17 of 21 comparisons	14 of 21 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (4/4)	85.7 (6/7)
	Kappa[c]	NA	NA	NA
	Precision	Most precise[d]	Least precise	Most precise[d]
Ladder	No result produced[a]	0 of 28 comparisons	25 of 28 comparisons	23 of 28 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (3/3)	100 (5/5)
	Kappa[c]	NA	NA	NA
	Precision	2nd most precise	Least precise	Most precise

[a]No result produced indicated the number of drug-drug comparisons for which the method was unable to produce a result either due to no studies with a common comparator or an insufficient number of studies with a common comparator.

[b]Percent agreement calculated only for drug-drug comparisons that both methods were able to produce a result. Results were considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for Bayesian MTC) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment.

[c]Kappa was calculated only for comparisons when both methods were able to produce a result. Not applicable (NA) indicates that the statistical program was unable to calculate a kappa because of insufficient data (because there were less than 2 levels).

[d]Neither the Bayesian MTC method or the Bucher method were clearly more precise than the other. Each of them reported the most precise results for about half of the drug-drug comparisons.

Abbreviations: MTC=mixed treatment comparison; NA=not applicable

Biologic Disease Modifying Antirheumatic Drugs for Rheumatoid Arthritis: ACR50 (Binary Outcome)

Appendix Tables D-12 through D-16 provide results of our analyses for each of the network patterns (sub-components of the full network) selected, including a placebo star, loop, one closed loop, and ladder for each drug-drug comparison for Bayesian MTC meta-analyses, frequentist meta-regression, the Bucher method, and frequentist logistic regression (Appendix D). Table 8 summarizes the number of comparisons for which each method was unable to produce a result as well as the percent agreement and kappa between each frequentist method and the Bayesian MTC method, by network pattern. On average, logistic regression produced the most precise results for all of the network patterns; the Bayesian MTC meta-analysis produced the third most precise or the least precise results, depending on the network pattern (Appendix Tables E-12 through E-16).

Table 8. Comparison of findings for ACR50 response between Bayesian MTC meta-analysis findings and those of frequentist methods for various network patterns

Network Pattern	Outcome	Bayesian MTC	Meta-Regression	Bucher Method	Logistic Regression
Star	No result produced[a]	0 of 28 comparisons	0 of 28 comparisons	0 of 28 comparisons	0 of 28 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (28/28)	89.3 (25/28)	78.6 (22/28)
	Kappa[c]	NA	1.000	0.7308	0.5333
	Precision	3rd most precise	Least precise	2nd most precise	Most precise
Loop	No result produced[a]	0 comparisons	0 of 1 comparisons	0 of 1 comparisons	0 of 1 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (1/1)	100 (1/1)	100 (1/1)
	Kappa[c]	NA	NA	NA	NA
	Precision	3rd most precise	Least precise	2nd most precise	Most precise
One closed loop with adalimumab	No result produced[a]	0 of 3 comparisons	0 of 3 comparisons	0 of 3 comparisons	0 of 3 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (3/3)	100 (3/3)	100 (3/3)
	Kappa[c]	NA	NA	NA	NA
	Precision	Least precise	3rd most precise	2nd most precise	Most precise
One closed loop with etanercept	No result produced[a]	0 of 3 comparisons	0 of 3 comparisons	0 of 3 comparisons	0 of 3 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (3/3)	100 (3/3)	100 (3/3)
	Kappa[c]	NA	1.000	1.000	1.000
	Precision	3rd most precise	Least precise	2nd most precise	Most precise
Ladder	No result produced[a]	0 of 3 comparisons	1 of 3 comparisons	1 of 3 comparisons	0 of 3 comparisons
	% agreement with Bayesian MTC[b]	NA	100 (2/2)	100 (2/2)	66.7 (2/32)
	Kappa[c]	NA	1.000	NA	NA
	Precision	Least precise	3rd most precise	2nd most precise	Most precise

[a]No result produced indicated the number of drug-drug comparisons for which the method was unable to produce a result either due to no studies with a common comparator or an insufficient number of studies with a common comparator.

[b]Percent agreement calculated only for drug-drug comparisons that both methods were able to produce a result. Results were considered to agree if both methods produced a non-statistically significant (for frequentist methods) or unimportant (for

Bayesian MTC) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment.

[c]Kappa was calculated only for comparisons when both methods were able to produce a result. Not applicable (NA) indicates that the statistical program was unable to calculate a kappa because of insufficient data (because there were less than 2 levels).

Comparison of Bayesian MTC Results for the Full Networks with Bayesian MTC Results for Sub-components

In this section, we present findings of the Bayesian MTC meta-analyses for each dataset for the full network and for sub-components of the network. Because the data differ for each scenario, the measures are not intended to be directly comparable for the full network and for the various sub-components of the network, but are presented here to show model fit under each scenario. The findings for the full network represent those for the complete literature on each set of medications (at the time of our CERs on each topic).[16, 17] Those for the sub-components examined hypothetical scenarios that would occur if the bodies of evidence were more limited.

Second-Generation Antidepressants: Binary Outcome

We ran MTC meta-analyses for the binary outcome of response to SGAs under five different network scenarios, the full network and four sub-component network patterns: star, loop, one closed loop, and ladder (as described in the Methods chapter). Network figures in Appendix C detail the geometry of these sub-components.

Measures of model fit are reported in Table 9. Estimates of between-study heterogeneity (tau squared) ranged from 0.12 to 0.44. Deviance information criterion (DIC) values varied across scenarios. When examining the total residual deviance for each scenario, all of them reasonably approximate the number of data points available, suggesting good model fit. The ladder and star pattern models most closely approximated the number of data points available.

Table 9. Measures of model fit for second-generation antidepressants for response

Statistic	Full Network	Star	Loop	One Closed Loop	Ladder
Total residual deviance (total number of datapoints available)	135.3 (140)	35.96 (36)	25.64 (30)	37.67 (42)	39.36 (40)
Deviance Information Criterion	259.45	246.67	188.42	271.67	270.31
Between-study heterogeneity	0.2571	0.4427	0.1172	0.1196	0.3019

One feature of the Bayesian framework for MTC is the ability to directly calculate the probability that each drug is the best treatment. Table 10 shows the results of this statistic for the five scenarios. Since not every treatment was available in each network pattern, some comparisons were not applicable for every network pattern.

Table 10. Probability of best treatment for second-generation antidepressants for response

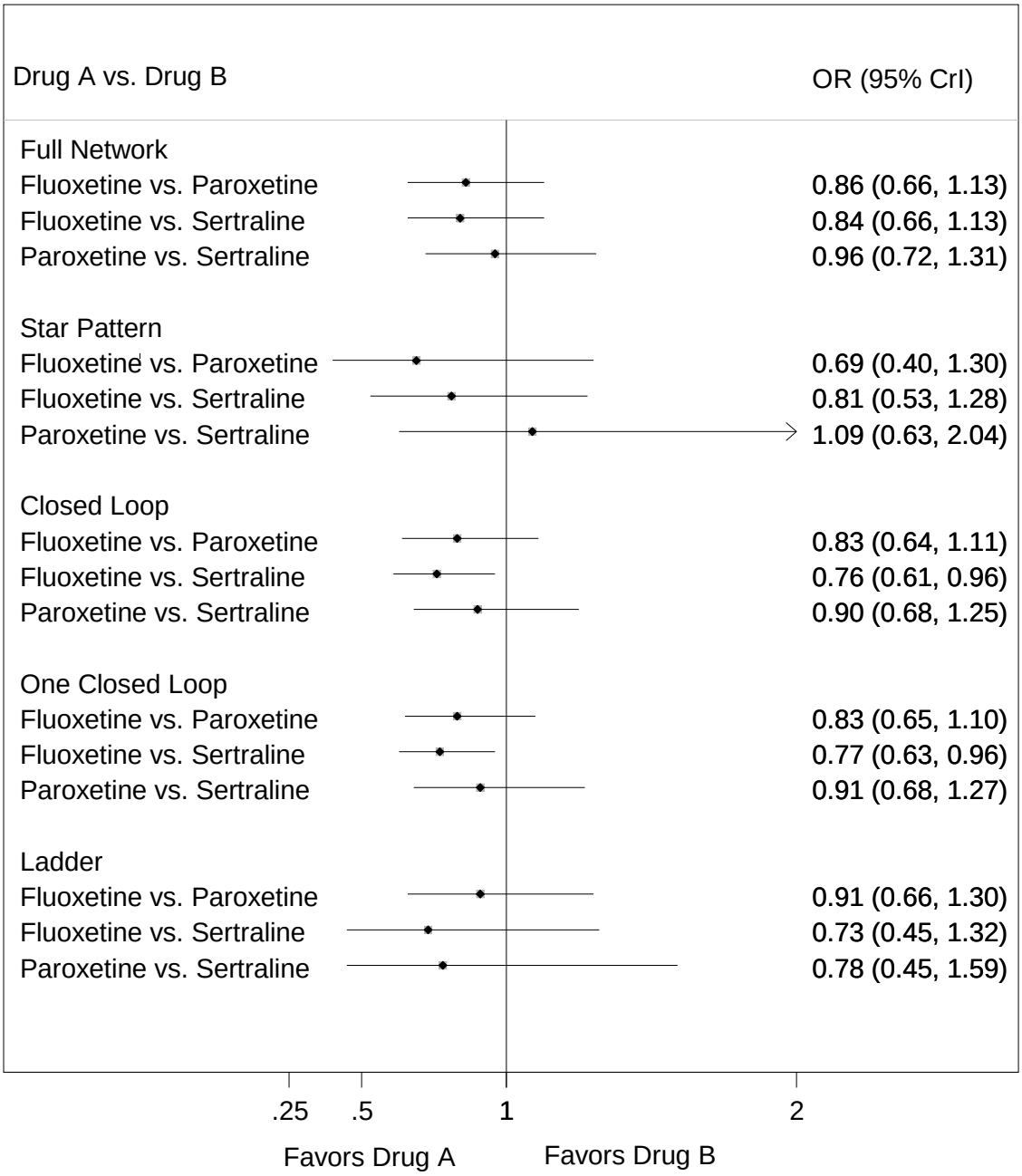
Statistic	Full Network	Star	Loop	One Closed Loop	Ladder
Bupropion	0.005	NA	NA	0.057	0.034
Citalopram	0.366	NA	NA	0.218	0.268
Desvenlafaxine	0.023	0.019	NA	NA	NA
Duloxetine	0.000	0.005	NA	NA	0.002
Escitalopram	0.090	NA	NA	0.056	0.110
Fluoxetine	0.000	0.185	0.019	0.003	0.004
Fluvoxamine	0.280	0.255	NA	0.202	0.262
Mirtazapine	0.158	NA	NA	0.077	0.061
Nefazodone	0.038	0.077	NA	NA	NA
Paroxetine	0.003	0.210	0.182	0.070	0.021
Sertraline	0.003	0.007	0.798	0.317	0.069
Trazodone	0.008	NA	NA	0.001	0.002
Venlafaxine	0.028	0.243	NA	NA	0.168

Abbreviations: NA=Not applicable.

For the full network, citalopram had the greatest probability of being the best treatment for achieving response (36.6%), followed by fluvoxamine. However, none of the treatments was particularly dominant to where we would have high confidence in that treatment truly having greater efficacy—as we might suggest if one reached a probability of 90 percent or more, for example. For the star network pattern, fluvoxamine had the greatest probability, followed by venlafaxine. But, again, none of the treatments was particularly dominant. The medication with the greatest probability of being the best treatment varied for each of the five scenarios. Only the closed loop had a fairly dominant medication—as sertraline had a probability of almost 80 percent for being the best treatment.

Given that not all of the comparisons were available in each scenario, we focused on the three drugs that were represented in each pattern: fluoxetine, paroxetine, and sertraline. Figure 5 shows the odds ratios and 95 percent credible intervals for the relative response of the antidepressants for each scenario. The full set of all pairwise comparisons is reported in Appendix D (Appendix Tables D-1, D-4, D-5, D-6, and D-7). Generally, the results show that the three drugs are not significantly different in odds of response. The two models with the least amount of connected data (star and ladder) also had the greatest heterogeneity and thus wider credible intervals around the mean. The response profiles for the loop and one closed loop mirrored the full network (with very similar point estimates), but found a significant difference between fluoxetine and sertraline, in part due to the lower between-studies heterogeneity of these reduced datasets.

Figure 5. Results of Bayesian MTC meta-analysis for five scenarios: odds ratio (95% credible interval) comparing fluoxetine, paroxetine, and sertraline for achieving response



Second-Generation Antidepressants: Continuous Outcome

For the second illustration with the antidepressant data, response is now represented on a continuous scale by using mean change from baseline in HAM-D in each treatment arm. We ran MTC meta-analyses using a normal likelihood and identify link function for each of the five different network scenarios. Network figures in Appendix C detail the geometry of these scenarios.

Out of the 78 pairwise comparisons available in the full network, there were 21 available in the star pattern, 3 in the loop, 21 in the one closed loop, and 28 in the ladder. The full results are reported in Appendix D (Tables D-2, D-8, D-9, D-10, and D-11). Although the point estimates vary across the patterns, likely due to the availability of direct head-to-head studies, the confidence intervals cross the line of no difference in all but one of the comparisons. As in the case of the binary outcome (response), there is little evidence from these data to support any important differences in the efficacy of the 13 antidepressants.

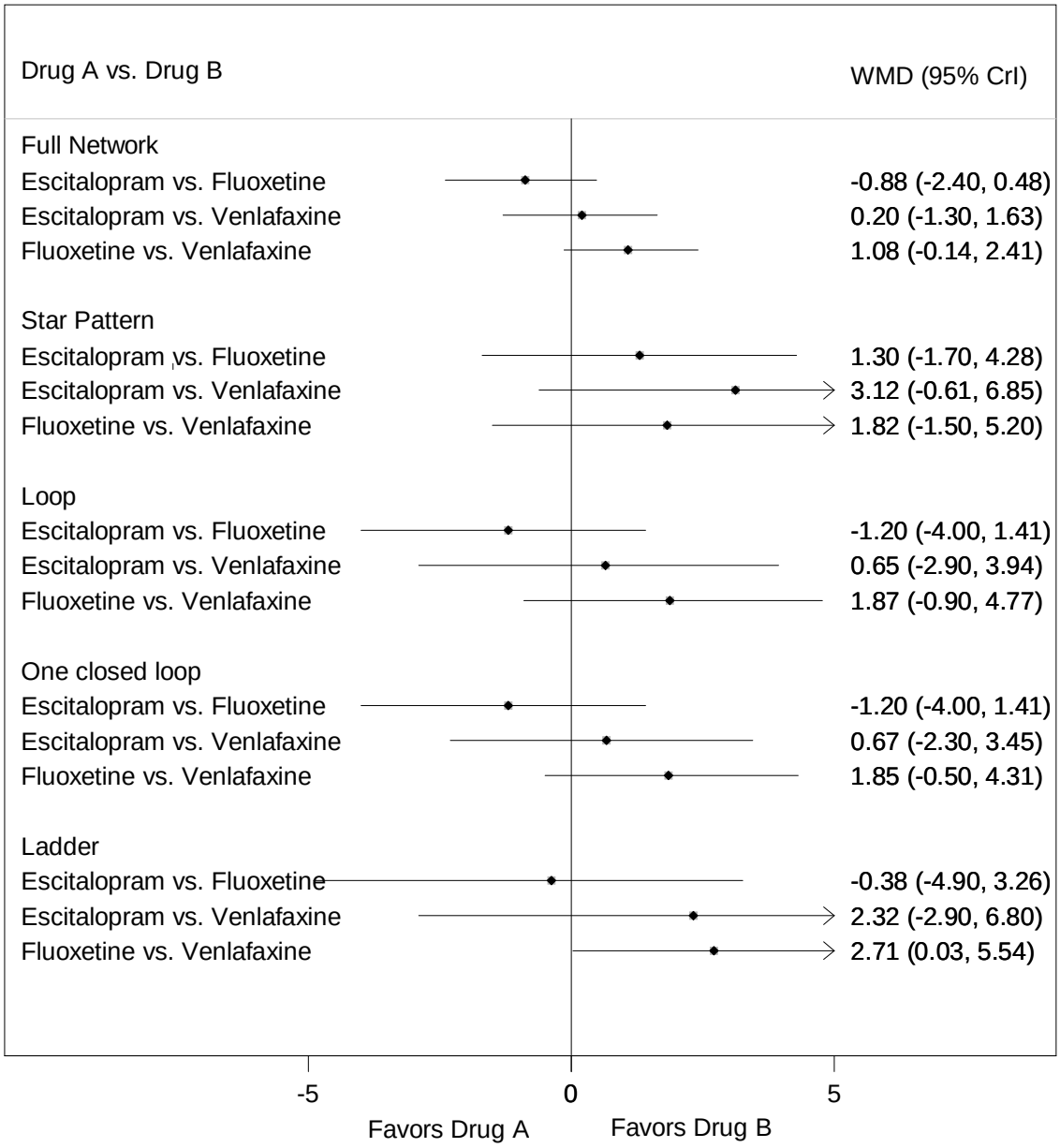
Measures of model fit are reported in Table 11. Estimates of between-study heterogeneity (tau squared) ranged from 0.86 to 1.97. For the total residual deviance for each scenario, the star, loop, and ladder most closely approximated the number of data points available, but fairly close approximations were found for all scenarios, suggesting good model fit.

Table 11. Measures of model fit for second-generation antidepressants for mean change from baseline in HAM-D

Statistic	Full Network	Star	Loop	One Closed Loop	Ladder
Total residual deviance (total number of datapoints available)	86.93 (80)	41.27 (40)	18.81 (18)	39.64 (38)	26.71 (26)
Deviance Information Criterion	254.08	120.76	59.866	122.637	75.126
Between-study heterogeneity	0.8564	0.8637	1.974	1.683	1.192

As in the example with a binary outcome, we looked closer at three pairwise comparisons available under each scenario (Figure 6). Generally, the results show that the three drugs are not significantly different in mean change from baseline in HAM-D.

Figure 6. Results of Bayesian MTC meta-analysis for five scenarios: weighted mean difference (95% credible interval) comparing escitalopram, fluoxetine, and venlafaxine for mean change from baseline in HAM-D



Biologic DMARDs for Rheumatoid Arthritis: Binary Outcome

For the rheumatoid arthritis dataset, the majority of the trials were placebo-controlled; only one trial included a head-to-head comparison, between adalimumab and infliximab. Therefore, the choice of sub-component network patterns was more limited than for the SGA dataset. We ran MTC meta-analyses on the binary outcome of ACR50 response for five different network scenarios, the full network and four sub-network patterns: star, loop, one closed loop, and ladder (as described in the Methods chapter). Network figures in Appendix C detail the geometry of these scenarios.

Measures of model fit are reported in Table 12. Estimates of between-study heterogeneity (tau squared) were very similar across scenarios (range 0.27 to 0.33). Total residual deviance data show that the full network and the loop most closely approximated the number of data points available, suggesting good model fit. The other scenarios had more sizeable differences between total residual deviance and the number of data points.

Table 12. Measures of model fit for biologic DMARDs for rheumatoid arthritis for achieving ACR50

Statistic	Full Network	Star	Loop	One Closed Loop	Ladder
Total residual deviance (total number of datapoints available)	65.54 (62)	62.29 (40)	17.01 (18)	22.87 (38)	16.03 (26)
Deviance Information Criterion	394.56	374.04	110.95	149.89	101.45
Between-study heterogeneity	0.3006	0.3193	0.3315	0.2668	0.3899

Table 13 shows the probability that each drug is the best treatment for achieving ACR 50 response for each of the five scenarios. Since not every treatment was available in each network pattern, the probabilities for some treatments were not reported (i.e., not applicable) for some network patterns. For all five scenarios, etanercept had the greatest probability of being the best treatment for achieving ACR50 response, with probabilities of 94.9 percent or higher.

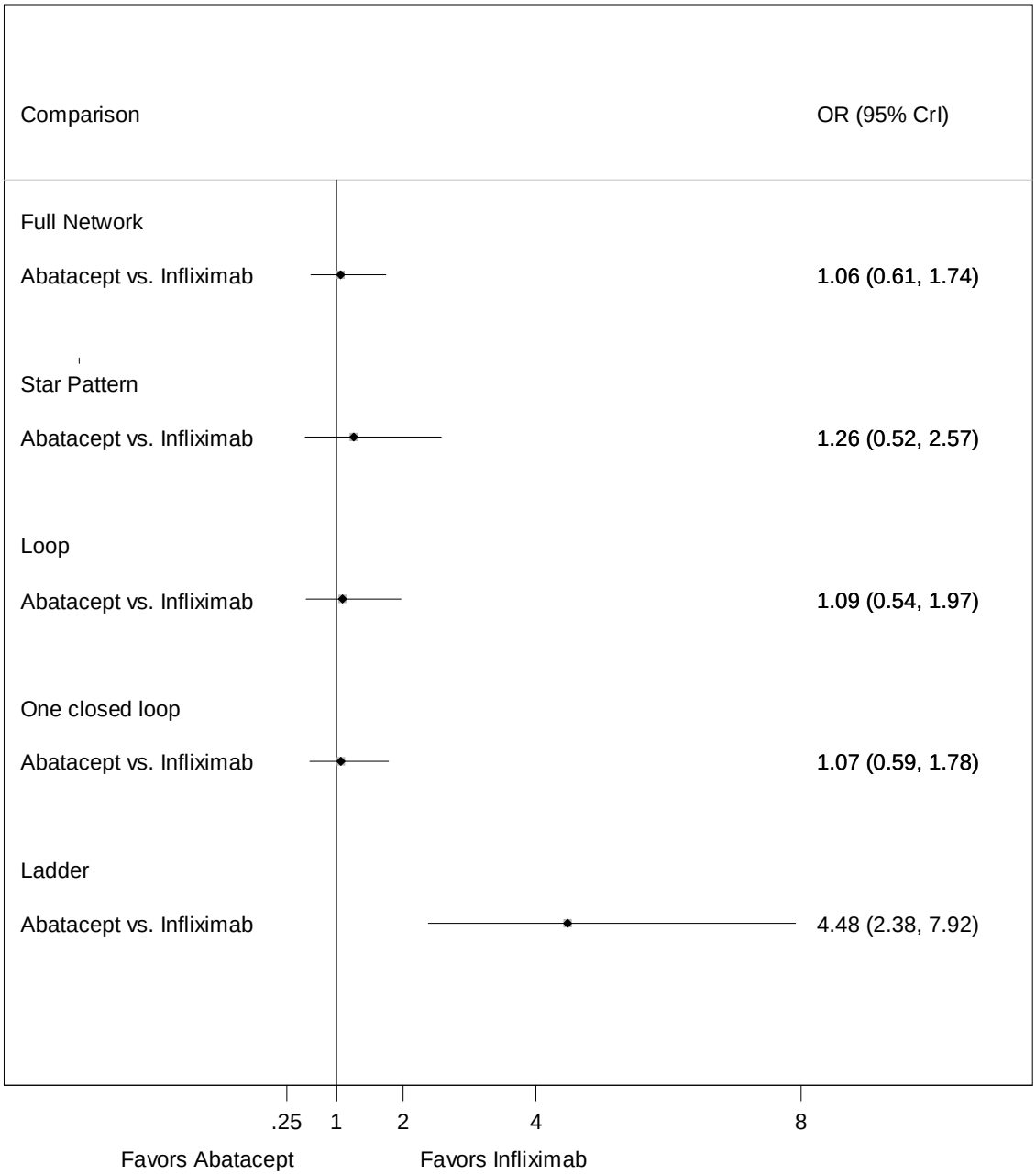
Table 13. Probability of best treatment for biologic DMARDs for rheumatoid arthritis for achieving ACR50

Statistic	Full Network	Star	Loop	One Closed Loop (ETA)	Ladder
Abatacept	0.000	0.001	0.4526	0.001	0.000
Adalimumab	0.003	0.003	NA	NA	NA
Anakinra	0.000	0.000	NA	NA	NA
Etanercept	0.953	0.949	NA	0.999	0.993
Golimumab	0.040	0.041	NA	NA	NA
Infliximab	0.000	0.001	0.5473	0.001	0.007
Rituximab	0.003	0.004	NA	NA	NA
Tocilizumab	0.000	0.000	NA	NA	NA

Abbreviations: NA=Not applicable; ETA=Etanercept.

Given that not all of the comparisons were available in each scenario, we focused on the two drugs that are represented in each pattern. Figure 7 shows the odds ratios and 95 percent credible intervals for the relative response for each scenario. The full set of all pairwise comparisons is reported in Appendix D (Appendix Tables D-3, D-12, D-13, D-14, D-15 and D-16). Generally, the results show that the two drugs in Figure 7 are not significantly different in odds of ACR50 response for four of the scenarios; the ladder found greater response for infliximab than for abatacept.

Figure 7. Results of Bayesian MTC meta-analysis for five scenarios: odds ratio (95% credible interval) comparing abatacept and infliximab for treatment response (ACR50)



Key Question 2. Meta-Regression with Bayesian MTC Meta-Analysis

This key question aims to illustrate the use of meta-regression as a technique for exploring sources of heterogeneity of treatment effects in CERs. There is often a need to explore heterogeneity in treatment effects in terms of another variable (e.g., an effect modifier), and many software packages facilitate meta-regression for pairwise meta-analysis. But, until

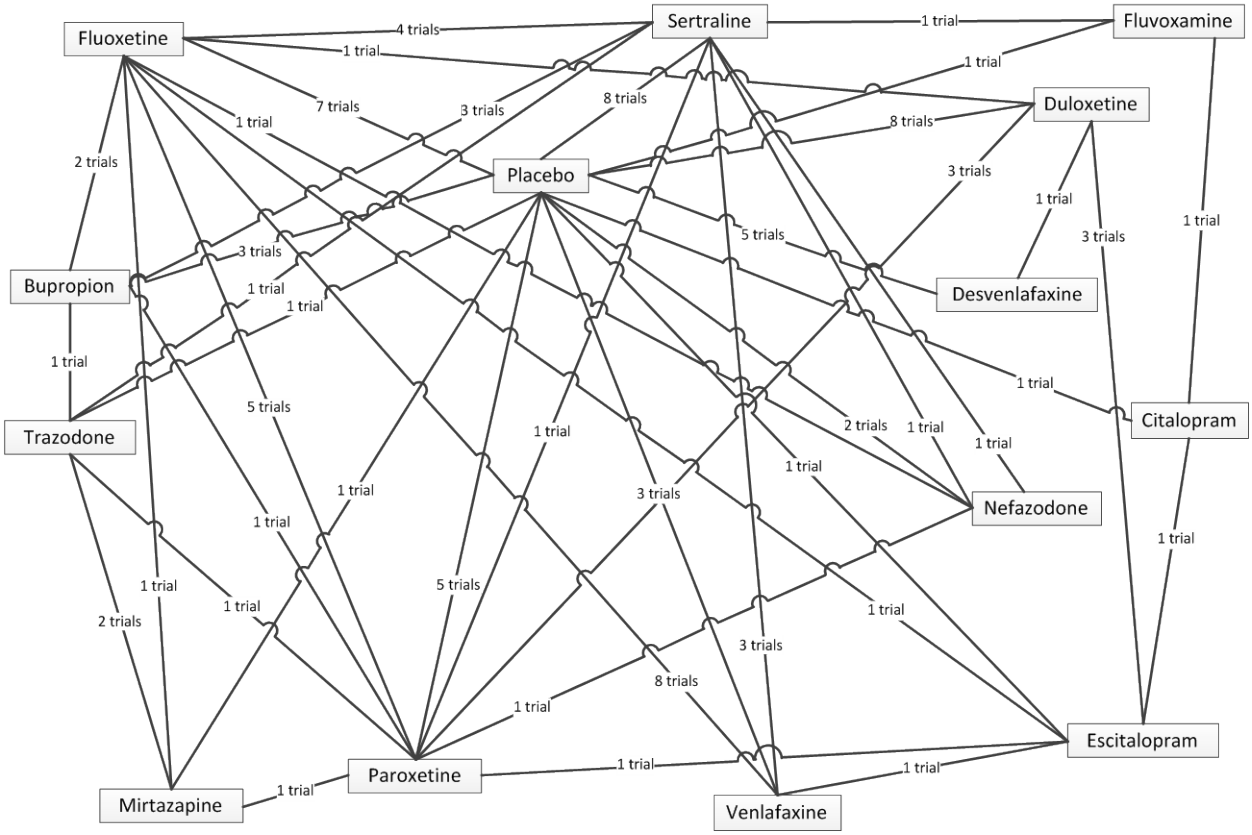
recently, there has not been a readily available and straightforward process for conducting meta-regression with Bayesian MTC in WinBUGS. The NICE Decision Support Unit recently published Technical Support Document 3 (TSD 3)[5] to illustrate the use of different types of meta-regression within Bayesian MTC. Bayesian MTC allows us to incorporate all of the available evidence into one analysis, which, combined with the ability to explore potential effect modifiers, can be an important analytic tool when conducting CERs.

We examine two types of meta-regression in this key question: one relates to exploring subgroup effects with a binary covariate and the other to exploring interaction effects with a continuous covariate. Our purpose here is to use meta-regression with Bayesian MTC meta-analysis within the context of two real-world scenarios from recent CERs and to examine two clinically important questions.

Meta-Regression with a Subgroup Indicator Covariate

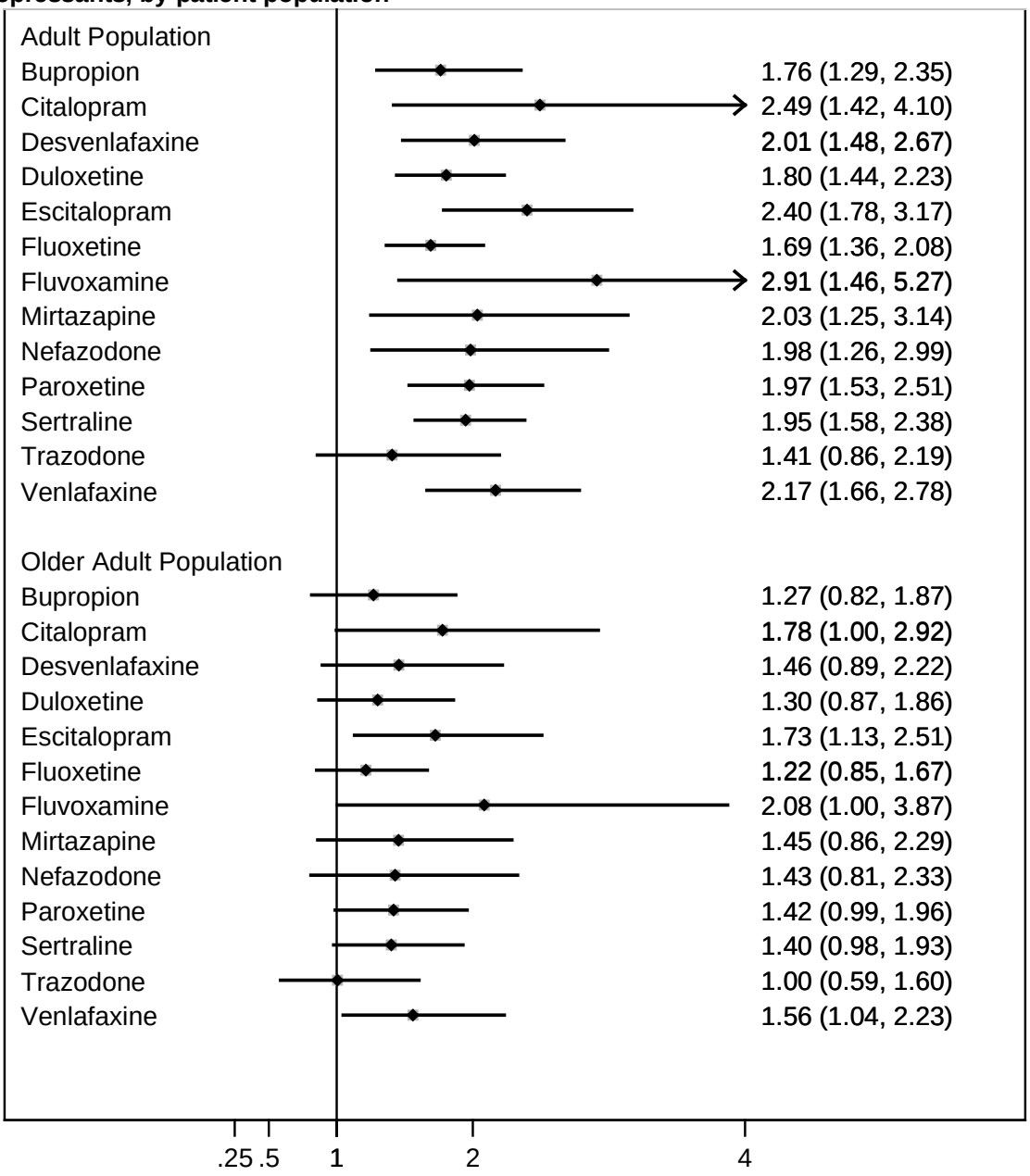
We hypothesized that differences in efficacy of SGAs may exist between older adults (≥55 years) and adults of any age. We used 72 trials of SGAs; 64 trials addressed efficacy in the general adult population and eight trials conducted exclusively in the older adult population. The outcome was treatment response as measured by 50 percent or greater improvement from baseline on the HAM-D. Figure 8 shows the evidence network for the 72 trials. We intended to test the interaction effect and compare the odds ratios for the treatments within each population.

Figure 8. Evidence network for subgroup meta-regression to assess whether efficacy of second-generation antidepressants differs for older adults



For this analysis, we used a random effects logistic regression model, including the appropriate adjustment for correlations within multi-arm trials. Convergence was checked via trace plots, and posterior means were calculated after 100,000 iterations following a burn-in of 20,000 iterations. We used a common (single) interaction term. Figure 9 shows the odds ratios of treatment response for each of the SGAs, relative to placebo, within each age subgroup.

Figure 9. Odds ratios (95% credible interval) of treatment response for second-generation antidepressants, by patient population



Within the general adult population, efficacy was supported for all antidepressants with the exception of trazodone. When estimating these effects for the older adult population within the

same meta-regression model, the efficacy for each drug appears to be diminished with only escitalopram and venlafaxine maintaining statistical superiority compared with placebo. This trend is supported by the interaction effect estimate of -0.34 (95% CrI, -0.696 to 0.006). This represents, on average, a -0.34 reduction in the log odds of response (not for the odds ratio scale). While approaching marginal statistical significance, the interaction estimate indicates a trend toward lower response rates in the older adult population. We compared the meta-regression model of 72 studies with the main adult analysis including 64 studies. The estimates for between-study heterogeneity were similar, although slightly reduced in the model containing the additional older adult studies (Table 14). The Deviance Information Criterion (DIC) increased slightly from 928.73 to 933.38. Both model estimates of total residual deviance were appropriate when compared with the number of data points available in each analysis, suggesting a good model fit in each case.

Table 14. Measures of model fit between main analysis and meta-regression

Statistic	Main Adult Analysis (n=64)	Meta-regression including Older Adults (n=72)
Total residual deviance (total number of datapoints available)	135.3 (140)	153.8 (158)
DIC	928.73	933.38
Between-study heterogeneity	0.2571	0.2506

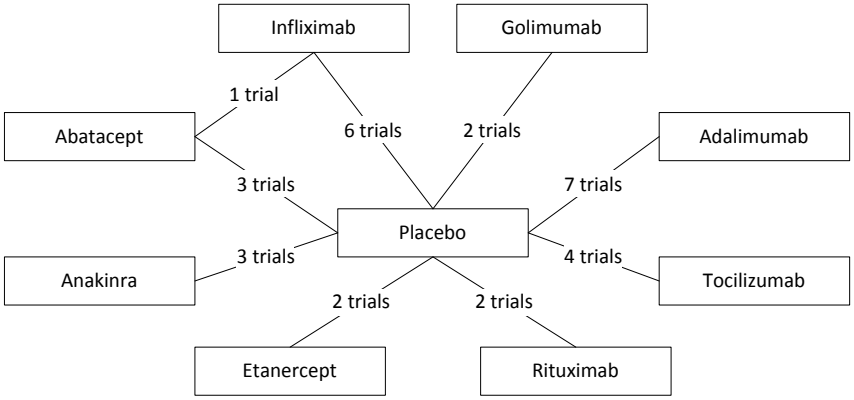
DIC=Deviance Information Criterion.

Meta-regression with a Continuous Covariate

We hypothesized that differences in efficacy of treatments for RA may exist for patients with varying durations of disease. Longer disease durations were hypothesized to be associated with higher levels of treatment response. We aimed to test the effects of this covariate (mean disease duration) on the findings from our recent CER—with our Bayesian MTC meta-analysis finding that etanercept resulted in greater treatment response than most other biologic DMARDs.

Twenty-eight of the 31 trials included in our Bayesian MTC meta-analysis reported data on mean disease duration (in years). Both trials that did not report data on mean disease duration compared etanercept with placebo. If disease duration was presented as only an arm-level mean, the average of the arms was taken for the trial-level mean. All eight biologic DMARDs were represented in the 28 trials (Figure 10). The mean trial-level disease duration ranged from 3.9 to 13 years.

Figure 10. Evidence network for continuous covariate meta-regression to assess whether efficacy of biologic DMARDs differs for patients with longer disease duration



A random effects logistic regression model was used, including the appropriate adjustment for correlations within multi-arm trials. The model was fitted assuming a common interaction effect for all treatments. Convergence was checked via trace plots, and posterior means were taken after 100,000 iterations following a burn-in of 20,000 iterations.

Table 15 shows the probability of best treatment with and without the meta-regression covariate. The interaction effect was found to be statistically important and supported the hypothesis of increased treatment response with increased disease duration. When including the effect of mean disease duration, the probability that etanercept was the best treatment dropped from 0.961 to 0.678, indicating that treatments with higher reported response rates, including etanercept, may have appeared to be more efficacious due to the inclusion of studies enrolling patient populations with longer disease durations.

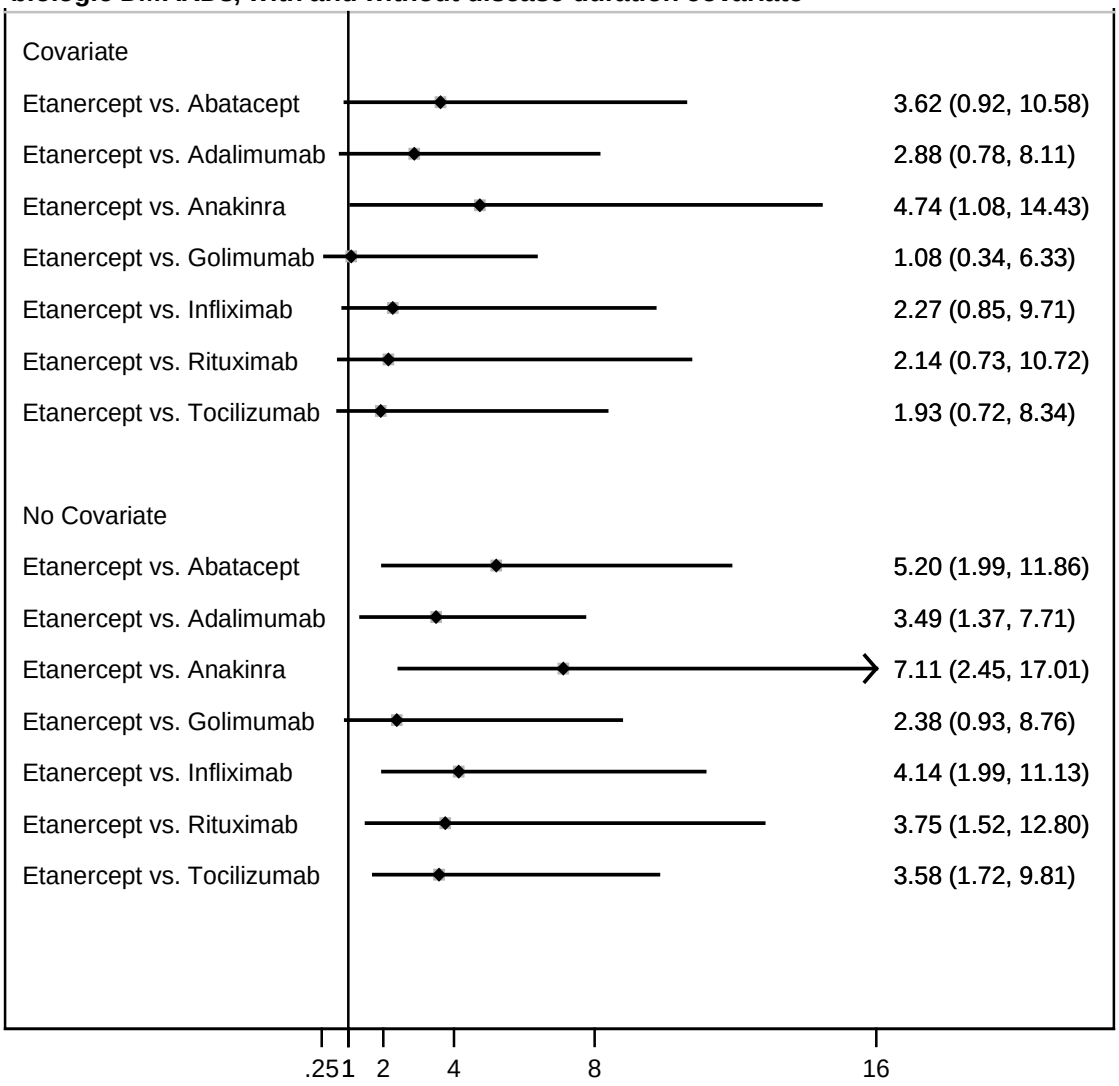
Table 15. Interaction effect and probability of best treatment

Probability of Best Treatment	No Covariate	With Disease Duration Covariate
Interaction Effect	NA	0.093 (95% CrI: 0.005 to 0.182)
Abatacept	0.001	0.001
Adalimumab	0.003	0.007
Anakinra	0.000	0.000
Etanercept	0.961	0.678
Golimumab	0.033	0.286
Infliximab	0.000	0.002
Rituximab	0.003	0.015
Tocilizumab	0.001	0.010

Abbreviations: CrI = credible interval; NA = Not applicable.

Figure 11 presents odds ratios for etanercept response compared with other biologic DMARDs from the model with and without the covariate. When examining the relative effects of etanercept compared with the other biologic DMARDs, we see that when controlling for years of disease severity, the relative efficacy of etanercept is reduced. While trending in the direction of greater efficacy compared with all of the other biologic DMARDs, statistical significance was only reached for the comparison with anakinra. However, even with the muted effect of etanercept in the model controlling for disease duration level, when looking at the probability of best treatment it appears the only other biologic with competing efficacy is golimumab. When predicting the estimated odds ratios for different disease durations, we found that odds ratios for etanercept compared with placebo ranged from 7.35 with one year of RA to 14.99 with 10 years of the disease.

Figure 11. Odds ratios (95% credible interval) of treatment response for Etanercept compared with other biologic DMARDs, with and without disease duration covariate



Key Question 3. Stability of the Bayesian MTC Analyses for Different Numbers of Studies and Network Pattern Assumptions

The motivation for this question arose from uncertainty about how the number of studies per comparison and different network scenarios interact with the ability of the model to produce valid results. We aimed to explore this question through an empirical study with simulated data. We anticipated the simulation study would provide insights about how many studies are needed in order for the Bayesian MTC meta-analysis to produce valid results. We first describe the results from our equivalent efficacy dataset and then describe those from our superior efficacy dataset.

Equivalent Efficacy Scenario

Table 16 shows the mean and standard deviation of the probability that each drug was the best in the 1,000 model runs for each network pattern and each available number of studies for the equivalent efficacy scenario. In the equivalent efficacy scenario, the pre-determined mean response (standard deviation) of drug one (placebo) was 0.10 (0.03); the pre-determined mean response (standard deviation) for each of drugs two, three, and four was 0.50 (0.05).

Table 16. Simulation results: Probability of best treatment under equivalent efficacy scenario

Network pattern	Number of Studies for Each Drug-Drug Comparison	Drug 1 (placebo) Mean (SD)	Drug 2 Mean (SD)	Drug 3 Mean (SD)	Drug 4 Mean (SD)
Star	1	0.01 (0.004)	0.34 (0.045)	0.33 (0.045)	0.32 (0.048)
	2	2×10^{-3} (2×10^{-4})	0.34 (0.115)	0.34 (0.112)	0.33 (0.116)
	3	1×10^{-6} (1×10^{-5})	0.33 (0.131)	0.33 (0.133)	0.34 (0.136)
	5	0.00 (0.000)	0.33 (0.144)	0.34 (0.141)	0.33 (0.139)
	10	0.00 (0.000)	0.33 (0.153)	0.34 (0.150)	0.33(0.150)
Loop	1	0.03 (0.009)	0.24 (0.054)	0.34 (0.049)	0.38 (0.040)
	2	5×10^{-4} (5×10^{-4})	0.27 (0.152)	0.33 (0.140)	0.40 (0.116)
	3	3×10^{-6} (2×10^{-5})	0.27 (0.174)	0.33 (0.168)	0.40 (0.132)
	5	0.00 (0.000)	0.27 (0.190)	0.33 (0.180)	0.40 (0.139)
	10	0.00 (0.000)	0.27 (0.201)	0.33 (0.190)	0.40 (0.150)
One Closed Loop	1	0.01 (0.004)	0.25 (0.077)	0.25 (0.077)	0.49 (0.004)
	2	1×10^{-4} (2×10^{-4})	0.26 (0.126)	0.25 (0.126)	0.49 (0.005)
	3	7×10^{-7} (1×10^{-5})	0.25 (0.137)	0.26 (0.137)	0.49 (0.005)
	5	0.00 (0.000)	0.25 (0.146)	0.26 (0.146)	0.49 (0.006)
	10	0.00 (0.000)	0.25 (0.148)	0.26 (0.148)	0.49 (0.006)
Ladder	1	0.11 (0.019)	0.30 (0.054)	0.24 (0.062)	0.35 (0.053)
	2	0.003 (0.002)	0.37 (0.169)	0.28 (0.186)	0.36 (0.168)
	3	4×10^{-5} (1×10^{-4})	0.37 (0.195)	0.28 (0.218)	0.35 (0.195)
	5	0.00 (0.000)	0.36 (0.209)	0.29 (0.234)	0.35 (0.203)
	10	0.00 (0.000)	0.36 (0.225)	0.30 (0.246)	0.34 (0.215)

Abbreviations: SD = standard deviation

In the equivalent efficacy scenario, we would expect to find that there are no differences between drugs two, three, and four, while drug one (placebo) should clearly be found the least efficacious. Table 17 shows the bias of the simulation results, which is the difference between observed and expected probabilities. For this data scenario, the bias ranged from 0.00 to 0.16, the latter being mainly in the one closed loop pattern. In the following bullets, we summarize our findings:

- For the star and ladder network patterns the correct conclusion was generally supported for each scenario (with varying numbers of studies for each drug-drug comparison). Even in cases where one drug had a slightly higher probability than expected, the difference was not sufficiently large to lead to the wrong conclusion (that one drug was superior to other drugs).
- The closed loop and one closed loop patterns had higher predicted means for the fourth drug in the network, even though drug four is not more efficacious than drugs two or three. The closed loop pattern found drug four to be the best drug in almost half of the iterations.

Table 17. Bias of Simulation Results: Difference between observed and expected probability of best treatment under equivalent efficacy scenario

Expected Value		**0.00**	**0.33**	**0.33**	**0.33**
Network pattern	Number of Studies for Each Drug-Drug Comparison	Drug 1 (placebo) Bias	Drug 2 Bias	Drug 3 Bias	Drug 4 Bias
Star	1	0.01	0.01	0.00	0.01
	2	0.00	0.01	0.01	0.00
	3	0.00	0.00	0.00	0.01
	5	0.00	0.00	0.01	0.00
	10	0.00	0.00	0.01	0.00
Loop	1	0.03	0.09	0.01	0.05
	2	0.00	0.06	0.00	0.07
	3	0.00	0.06	0.00	0.07
	5	0.00	0.06	0.00	0.07
	10	0.00	0.06	0.00	0.07
One Closed Loop	1	0.01	0.08	0.08	0.16
	2	0.00	0.07	0.08	0.16
	3	0.00	0.08	0.07	0.16
	5	0.00	0.08	0.07	0.16
	10	0.00	0.08	0.07	0.16
Ladder	1	0.11	0.03	0.09	0.02
	2	0.00	0.04	0.05	0.02
	3	0.00	0.04	0.05	0.02
	5	0.00	0.03	0.04	0.02
	10	0.00	0.03	0.03	0.01

- Findings of analyses when one study was available for each comparison were generally very similar to findings when more studies (two, three, five, or ten) were available. Differences between results with one study available and those with ten studies available were often 0.01 or less and were never greater than 0.06 (for drugs two and three in the ladder pattern).

Superior Efficacy Scenario

Table 18 shows the mean and standard deviation of the probability that each drug was the best in the 1,000 model runs for each network pattern and each available number of studies for the superior efficacy scenario. In the superior efficacy scenario, the pre-determined mean response (standard deviation) for each of drugs one, two, and three was 0.20 (0.04); the pre-determined mean response (standard deviation) of drug four was 0.80 (0.04).

Table 18. Simulation results: Probability of best treatment under superior efficacy scenario

Network pattern	Number of Studies for Each Drug-Drug Comparison	Drug 1 Mean (SD)	Drug 2 Mean (SD)	Drug 3 Mean (SD)	Drug 4 Mean (SD)
Star	1	0.03 (0.009)	0.13 (0.016)	0.13 (0.017)	0.71 (0.023)
	2	8×10^{-4} (8×10^{-3})	0.01 (0.006)	0.010 (0.006)	0.98 (0.012)
	3	2×10^{-5} (5×10^{-4})	2×10^{-4} (5×10^{-4})	0.0002 (0.0005)	0.9995 (0.001)
	5	0.00 (0.000)	4×10^{-7} (7×10^{-6})	9×10^{-7} (9×10^{-6})	0.9999 (0.00001)
	10	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	1.00 (0.00)
Loop	1	0.05 (0.011)	0.08 (0.016)	0.16 (0.020)	0.71 (0.025)
	2	0.001 (0.001)	0.004 (0.004)	0.01 (0.010)	0.98 (0.014)
	3	2×10^{-5} (7×10^{-5})	0.0001 (0.0003)	0.0005 (0.001)	0.999 (0.001)
	5	0.00 (0.000)	0.00 (0.000)	1.4×10^{-6} (1×10^{-5})	0.99999 (1×10^{-5})
	10	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	1.00 (0.00)
One Closed Loop	1	0.17 (0.076)	0.17 (0.079)	0.17 (0.077)	0.49 (0.003)
	2	0.17 (0.118)	0.17 (0.119)	0.17 (0.120)	0.50 (0.005)
	3	0.16 (0.129)	0.17 (0.131)	0.17 (0.129)	0.50 (0.005)
	5	0.16 (0.132)	0.17 (0.133)	0.16 (0.131)	0.50 (0.005)
	10	0.17 (0.143)	0.17 (0.137)	0.16 (0.136)	0.50 (0.007)
Ladder	1	0.17 (0.023)	0.09 (0.018)	0.05 (0.11)	0.69 (0.030)
	2	0.01 (0.009)	0.004 (0.004)	0.001 (0.001)	0.98 (0.013)
	3	6×10^{-4} (0.001)	1×10^{-4} (0.0003)	2×10^{-5} (7×10^{-5})	0.999 (0.001)
	5	1×10^{-6} (1×10^{-5})	1×10^{-7} (3×10^{-6})	0.00 (0.000)	1.0 (0.00)
	10	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	1.0 (0.00)

Abbreviations: SD = standard deviation.

In the superior data scenario, we would expect to find no differences between drugs one, two, and three, while drug four should clearly be found to be the best treatment. Table 19 shows the bias of the simulation results for this data scenario. The differences between observed and expected probabilities ranged from 0.00 to 0.51, again with the bias in the one closed loop pattern being the most pronounced. In the following bullets, we summarize our findings:

- For each scenario (with varying numbers of studies and network patterns), the correct conclusion was generally supported, with one notable exception. In the one closed loop pattern, the model failed to definitively find drug four to be the most efficacious.
- Generally, there were no significant differences between the estimates generated with two, three, five, or ten studies available for each comparison.
- For the scenarios with one study available per comparison, although probabilities of best treatment differed numerically compared with those scenarios with more studies available, the findings were still indicative of superior efficacy of drug four. For drug four, differences between estimates (of the probability of being the best treatment) from analyses with one study available per comparison and those from analyses with two studies available per comparison ranged up to 0.29.

The estimates from Tables 16 and 18 are also presented graphically, as histograms in Appendix F.

Table 19. Bias of simulation results: Difference between observed and expected probability of best treatment under superior efficacy scenario

Expected Value		**0.00**	**0.00**	**0.00**	**1.00**
Network pattern	Number of Studies for Each Drug-Drug Comparison	Drug 1 (placebo) Bias	Drug 2 Bias	Drug 3 Bias	Drug 4 Bias
Star	1	0.03	0.13	0.13	0.29
	2	0.00	0.01	0.01	0.02
	3	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00
Loop	1	0.05	0.08	0.16	0.29
	2	0.00	0.00	0.01	0.02
	3	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00
One Closed Loop	1	0.17	0.17	0.17	0.51
	2	0.17	0.17	0.17	0.50
	3	0.16	0.17	0.17	0.50
	5	0.16	0.17	0.16	0.50
	10	0.17	0.17	0.16	0.50
Ladder	1	0.17	0.09	0.05	0.31
	2	0.01	0.00	0.00	0.02
	3	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	10	0.00	0.00	0.00	0.00

Discussion

Main Findings by Key Question

In this report, we addressed three Key Questions (KQs) using real-world bodies of literature (for KQs 1 and 2) from recent comparative effectiveness reviews (CERs) of second-generation antidepressants (SGAs) and treatments for rheumatoid arthritis (RA) and using simulated data (for KQ 3). Below, we summarize the main findings by KQ. We then address the implications, limitations, research gaps, and conclusions.

Key Question 1: Bayesian Mixed Treatment Comparison (MTC) Methods Compared With Several Frequentist Indirect Methods, and Performance for Different Types of Evidence Network Patterns

We compared the results of the Bayesian MTC approach with results of several frequentist analytic methods—specifically, frequentist meta-regression, the Bucher method (adjusted indirect comparisons), and frequentist logistic regression. Of note, the frequentist methods used are not the analogues of the Bayesian methods implemented; i.e., we did not compare findings with a frequentist approach to MTC. Our choice of methods to compare, and our manner of conducting the analyses, was based on our judgment regarding the methods most commonly used by analysts conducting CERs. Our results for these comparisons, such as those related to precision, are not necessarily generalizable to other datasets (e.g., comparative effectiveness of other medication classes).

We found one of the main differences between Bayesian MTC meta-analysis and the typically-applied approach for frequentist meta-regression and the Bucher method to be that the Bayesian MTC approach was able to calculate a result for all drug-drug comparisons of interest whereas the other methods were unable to produce a result for many comparisons of interest. This was not surprising; it is because Bayesian MTC meta-analysis is able to produce a result for all comparisons in a connected network, whereas we only calculated results for these other frequentist indirect methods when there was a common comparator. Our results showed that some frequentist methods were unable to compute results for *substantial* proportions of the drug-drug comparisons of interest for some of the datasets. Despite that, our results may actually underestimate the proportion that one might determine for inability to calculate a result if running these analyses for a real-world CER because some guidance for certain methods would suggest not even attempting some of the analyses without a certain minimum number of studies. For example, we conducted meta-regression for all comparisons with a direct common comparator for which the statistical program was able to calculate a result; however, some research indicates that a certain minimal number of studies is needed for the meta-regression to produce a reliable result.[29, 39] In other words, although technically we could produce results for meta-regression and the Bucher method for all of the comparisons for which we reported results, many of those, perhaps, should not be calculated if conducting a CER for the purpose of informing decisionmaking.

We were able to calculate results for all methods (Bayesian MTC and the three frequentist methods) for the RA dataset, but not for the SGA dataset. This is not surprising given the geometry of the two networks. The RA network is largely a star pattern with all treatments connected via placebo and with only one head-to-head study. The data is thus set up to allow any

of the methods to determine results for each drug-drug comparison of interest (because of the common placebo comparator for each drug-drug comparison of interest). The SGA network, however, is a much more complex network. In this complex network, there were often too few studies with a common comparator to allow some frequentist methods to calculate results for many comparisons of interest.

Regarding various network patterns, the frequentist meta-regression and the Bucher method (in the manner we applied them) were least likely to be able to compute results for comparisons of interest for ladder patterns and most likely to be able to compute results for star patterns. The geometry of ladder patterns, by definition, does not include common comparators for most comparisons of interest, as each treatment will be linked to two other treatments at most.

In the cases of comparisons for which we were able to produce results, the majority of these results were in general agreement (i.e., they all found no statistically significant difference [for frequentist analyses] or no important difference [for Bayesian analyses] or they found a significant or important difference favoring the same drug). However, for each of the full networks, and for each outcome of interest, there were instances when some of the frequentist methods produced results that did not agree with those of the Bayesian MTC analysis. In each case, we do not know for certain which of the results represents the truth; instead, we just know that the findings do not agree. One might presume that the Bayesian MTC analysis is more likely to approximate the truth because it is able to incorporate more information from the full dataset; however, this assumption has not been proven. Validating such a conclusion would require a large number of real-world examples or data simulations using numerous situations and assumptions.

When considering precision, we speculated whether the Bayesian MTC method might generally produce more precise results because of its ability to incorporate all data (from both direct head-to-head trials and placebo-controlled trials). Our results did not find this to be the case. Across all the different network patterns (including the full networks), the logistic regression method produced the most precise result for a greater number of drug-drug comparisons than any of the other methods, followed by the Bayesian MTC method, and then the Bucher method. The meta-regression method fairly consistently produced the least precise result. The differences in precision between the various analyses were sometimes very small (on the order of hundredths), but were larger in some cases (e.g., on the order of 2 to 5 points different for an odds ratio), and very large in rare cases (e.g., a difference of 20 to 50).

For some specific networks (five of the 16 networks or sub-component networks that we analyzed), the Bayesian MTC results were the most precise. From our findings, we are unable to determine if any particular network geometry makes the Bayesian MTC method more or less likely to produce the most precise results.

Key Question 2: Meta-Regression with Bayesian MTC Meta-Analysis

When we began this project, we were unaware of any available guidance on conducting meta-regression (i.e., adding study level covariates) with Bayesian MTC meta-analysis. We hoped to explore how this could be done and to write WinBUGS code to conduct such analyses for two scenarios. Subsequently, the NICE Decision Support Unit released a Technical Support Document (TSD) detailing use of the use of different types of meta-regression with Bayesian MTC meta-analysis.[5] Therefore, we followed their published guidance to conduct two meta-regressions—one to explore subgroup effects with a binary covariate and the other to explore

interaction effects with a continuous covariate. In each case, covariates were added to the existing logistic regression models used in KQ1.

For the first meta-regression, we hypothesized that differences in efficacy of SGAs may exist for older adults (≥55 years) than for younger adults. We explored this hypothesis through a subgroup meta-regression. We found a trend toward lack of efficacy or lesser efficacy for all of the SGAs in older adults. However, this finding could be due to case mix, as studies of older adults may enroll subjects with less severe depression at baseline.

For the second meta-regression, we hypothesized that treatment response may differ for patients with longer disease duration. We speculated that longer disease duration was related to more severe RA (and potentially more room for improvement). We explored this hypothesis through a continuous covariate meta-regression with mean disease duration as the continuous covariate. We focused on the response to etanercept to determine whether controlling for this covariate might alter findings because our MTC meta-analysis showed etanercept yielding a greater treatment response than most other biologic DMARDs. We found a trend toward greater efficacy for those with greater mean disease duration. In addition, we found that the superior relative efficacy of etanercept compared with other biologic DMARDs was reduced; furthermore, it was no longer significantly greater (based on the 95% credible intervals) than most of the other biologic DMARDs (with the exception of anakinra). However, reporting bias might underlie this finding because the only two trials (out of 31) that were not included in the analysis (because they did not report mean disease duration) compared etanercept and placebo.

Key Question 3: Stability of the Bayesian MTC Analyses for Different Numbers of Studies and Network Pattern Assumptions

Our simulations for KQ 3 revealed several important findings. First, the simulations validated the ability of the Bayesian MTC meta-analyses to produce results that would yield conclusions that reflect the truth for two scenarios—one with three medications of equivalent efficacy and a placebo arm and one with one medication of superior efficacy. The analyses were generally able to produce findings reflecting the underlying truth with only one study per comparison.

Second, for the equivalent efficacy scenario, results changed very little based on the number of studies available for each comparison. For the superior efficacy scenario, however, we found larger differences between scenarios with one available study for each comparison and scenarios with two studies for each comparison, and similar results with little variation between two and 10 studies per comparison.

Third, networks with a one closed loop pattern did not perform as expected and often produced results that might yield inaccurate conclusions. These networks produced results for the equivalent efficacy scenario that might lead investigators to conclude one of the drugs (drug four) to have the greatest likelihood of efficacy, regardless of the number of studies available for each comparison (i.e., the accuracy of results did not improve even with 10 studies per comparison). In contrast, for the superior efficacy scenario, the one closed loop found the correct drug to have the greatest likelihood of efficacy, but with much lower probability compared with the other network patterns (approximately 50% vs. well over 90%), regardless of the number of studies per comparison. We are uncertain as to why this occurred.

Implications for Comparative Effectiveness Reviews

Investigators have a number of analytic strategies from which to choose when conducting indirect comparisons (Bayesian MTC meta-analysis, frequentist meta-regression, the Bucher method, logistic regression, and others). Our findings indicate that the choice of method can have an important impact on the results. Although the various methods we compared using our two real-world datasets found results that generally agreed for the majority of drug-drug comparisons, we also found some instances where results differed for the same comparison. Even though we do not know with certainty which method is correct from these real-world comparisons, our findings identified several advantages of the Bayesian MTC approach (Table 20).

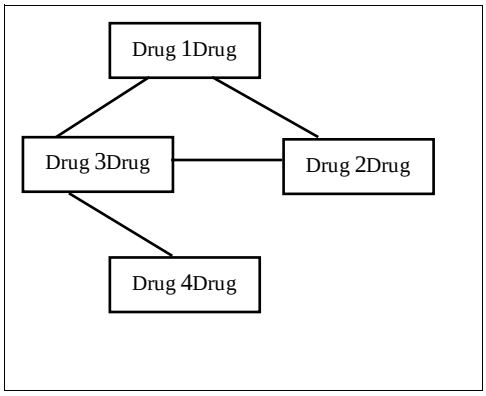
Table 20. Advantages and disadvantages of the Bayesian MTC approach

Advantages	Disadvantages
Able to incorporate both direct evidence (from head-to-head trials) and indirect evidence (e.g., placebo-controlled trials) into a single analysis	Less accessible to many investigators because the analyses require software that is unfamiliar or less familiar to many of them (usually run using WinBUGS)
Able to produce results for all comparisons of interest within a connected network (even for ladder network patterns or complex networks that limit the ability of other methods to get any results for some comparisons of interest)	Requires greater statistical expertise than some other methods
Able to directly calculate the probability that each drug is the best treatment	Further research is needed to evaluate performance in specific scenarios[a]
Able to adjust for correlations within multi-arm trials	Might not produce accurate results for one closed loop networks
Able to incorporate meta-regression to assess heterogeneity (e.g., for subgroups or to control for covariates), all within one model	Possibly sensitive to the prior probabilities chosen (therefore recommended to generally use flat priors)[40]
Appears to produce valid, accurate results for star and ladder network patterns	

[a]Such as those with various numbers of studies available per comparison, various sample sizes of studies, and complex network patterns.

Our simulations (KQ 3) supported the validity of the Bayesian MTC method for star and ladder network patterns. However, they raised some concerns about the validity of the Bayesian MTC method for one closed loop networks and possibly for loop patterns (albeit lesser concerns), as results did not converge on the pre-determined truth. These findings demonstrate the need for additional exploration of more hypothetical scenarios, and analysts should be cognizant of the fact that Bayesian MTC methods may not produce accurate results for one closed loop networks such as the one used in our simulation (Figure 12). We are uncertain as to why the results did not converge on the truth; we wonder whether a "dangling" treatment (where 1 treatment is peripheral to the rest of the network, like drug 4 in the figure below) might reduce accuracy.

Figure 12. One closed loop network pattern used in simulation



Surprisingly, our simulations did not find much difference between scenarios when one study was available for each comparison and those when more studies (two, three, five, or ten) were available. We initially expected that results from analyses with one study available for each

comparison would be much less likely to approximate the underlying truth, but found that this was not the case. We previously wondered whether Bayesian MTC meta-analyses should not be attempted when there is only one study per comparison, hypothesizing that results would be much less accurate. However, this hypothesis was not supported by our simulations.

Limitations

Our findings from KQs 1 and 2 are not necessarily generalizable to treatment networks for other medications or diseases. Some of the findings may be a consequence of the underlying body of evidence (e.g., the number of studies for each comparison, the variation of findings in studies making the same comparison) and all of its inherent biases (e.g., selective outcome reporting, publication bias) rather than the network pattern, per se.

In addition, our selection of various sub-component network patterns was based on maximizing the amount of data available for each sub-network. Due to limited time and resources, we could not evaluate every possible network pattern within each dataset. Other approaches to choosing network patterns could also have been informative (e.g., based on entry of drugs into the market, or chronologically by publication date). But, for the purposes of this report we felt that an approach based on maximizing the amount of data available was reasonable. Due to the nature of the sub-component network patterns (i.e., that they include only some of the available evidence from a body of literature), the findings for comparisons of sub-components of the full networks are for exploration of the various types of analyses only and should not be used to inform clinical decisions.

In KQ 1, we reported several measures to compare the findings of the Bayesian MTC method with various frequentist methods. The measures have some limitations. For example, regarding the *number of comparisons* for which each frequentist method was *unable to produce a result*, our results likely underestimate the numbers that some analysts might determine if running these analyses for a CER—because some guidance would suggest not even attempting some analyses without a certain minimum number of studies. For example, we conducted meta-regression for all comparisons that the statistical program was able to calculate a result, but it has been suggested that a certain minimal number of studies per covariate is needed for the meta-regression to produce a reliable result.[29, 39] In other words, the gap would actually be larger between the Bayesian MTC and meta-regression for the number of comparisons the method was unable to produce a result.

Another measure we used was the percent agreement between results of the Bayesian MTC method and those of each frequentist method. We considered results to agree if both methods produced a non-statistically significant (for frequentist analyses) or an unimportant (for Bayesian analyses) result for the comparison or if both analyses found a statistically significant or important result favoring the same treatment. This is an oversimplification of how results would be interpreted in a CER. For example, we did not determine whether the ultimate conclusions about comparative effectiveness or the strength of evidence grades would agree.

Our choice of analytic methods to compare was based on our judgment regarding the methods most commonly considered by analysts conducting CERs. In addition, we selected frequentist methods with some evidence to support their validity. However, we did not compare findings with the frequentist network meta-analysis method (i.e., Lumley method).[20] Our experience indicates that it is much more rarely used than the other methods, and it is generally felt that the Bayesian MTC approach has several advantages over frequentist network meta-analysis.

For Bayesian MTC analyses, we did not explore sensitivity analyses with informative priors. However, we do not believe that the current state of the literature would support using informative priors. In addition, we did not include inconsistency models to test model assumptions, or explore other methods in checking model convergence, such as the Gelman-Rubin statistic.

For KQ 2, both of our meta-regressions rely on averages taken over patients in the trials. As such, ecological bias is a potential limitation.

Our simulations for KQ 3 were limited by the assumptions that we made to develop the scenarios. Our simulations did not examine many scenarios observed in real-world networks. For example, we set each scenario to include an equal number of studies for each drug-drug comparison and we set the sample size at 100 subjects for each study. In addition, we used a non-random network structure, and we did not simulate incoherence. Findings might differ if one investigated network patterns with varying numbers of studies for different comparisons or with larger or smaller (or varying sample sizes), or if one incorporated additional complexity in the simulations.

For our simulations in KQ 3, we chose to output the probability that each drug was the best (i.e., most efficacious) for ease of presentation and to use an outcome recognizable to analysts familiar with Bayesian MTC. Being able to produce probability rankings is an advantage of the Bayesian MTC approach. However, one could argue that using mean treatment response would have been more appropriate, because it would have allowed outputs that were directly comparable with the underlying truths that we specified.

Finally, data were too sparse to include some analyses we had aimed to include at the outset. We wanted to run analyses for a continuous outcome from the RA dataset (mean change in Health Assessment Questionnaire score), as we had done for the SGA dataset, but too few studies reported sufficient data.

Future Research

We identified several issues that future research could address (Table 21). Many of these issues are related to further exploring the findings and limitations mentioned above. In addition to these, future research could focus on writing programs to conduct Bayesian MTC meta-analyses for programs that are more familiar to a greater number of investigators.

Conclusions

Bayesian MTC methods allow investigators to calculate results for many more comparisons of interest for some network patterns, including ladders and complex networks, than frequentist meta-regression or the Bucher method, in the manner that they are typically applied. When Bayesian MTC methods and various frequentist methods are each able to calculate results (as typically applied), the findings are usually in general agreement. However, findings differ for a small proportion (less than 10%) of comparisons, which could lead to differences in conclusions. Regarding precision, logistic regression produced the most precise result for more drug-drug comparisons than any of the other methods, followed by the Bayesian MTC method.

Our simulations support the validity of Bayesian MTC methods for star and ladder network patterns, but raise some concerns about one closed loop network patterns, and possibly about loop patterns (but less concern). Simulations generally found similar results for scenarios when one study was available for each comparison and those when more studies (two, three, five, or ten) were available.

Table 21. Possible targets for future research, by Key Question (KQ)

KQ	Potential Future Research
1	Our findings were based on just two real-world datasets. To determine how well our findings hold up for various real-world situations, it will be important to conduct similar analyses for other existing networks. Many (perhaps 50) publications have used Bayesian MTC meta-analyses; those could be explored.
1	For all of our Bayesian MTC meta-analyses, we ran a certain number of simulations (20,000) that were discarded and then used an additional 100,000 simulations in estimating the posterior probabilities. The impact of varying the number of simulations on the resulting findings is uncertain. Future research could address this impact to inform the most appropriate number of simulations.
1	There were numerous options of various network patterns we could have selected. Future research could use an approach similar to cumulative meta-analysis and explore how the evidence evolved over time thus, choosing network patterns based on the chronology of study publication.
1	Our analyses compared the results of various analytic approaches for both continuous and dichotomous outcomes, but we did not explore competing risk outcomes.[41] Future research could include competing risk outcomes.
2	We found a trend toward lack of efficacy or less efficacy for all of the SGAs in older adults. However, this finding could possibly be due to case mix, as studies of older adults may enroll subjects with less severe depression at baseline. Future analyses could explore this.
2	We found a trend toward greater efficacy for those with greater mean disease duration. However, this finding could possibly be due to case mix, or other factors that we did not consider. Future analyses could explore this further.
2	For our meta-regression models exploring (1) subgroup effects with a binary covariate and (2) interaction effects with a continuous covariate, we used models with the same interaction effect for all treatments. Other models have been described with independent, treatment-specific interactions and exchangeable, related treatment-specific interactions.[5] Future research could explore whether these other models would significantly alter findings.
3	For our simulations, we assumed a sample size of 100 subjects per study. Findings might differ if one investigated network patterns with larger or smaller (or varying sample sizes).
3	For our simulations, we assumed that each network had an equal number of studies available for each comparison of interest. Findings might differ if one investigated network patterns with varying numbers of studies for different comparisons.
3	For our simulations, we used four network patterns (star, loop, one closed loop, and ladder). Many additional variations of these patterns could be assessed to determine if findings differ. In addition, complex networks that resemble real-world datasets could be developed to attempt to validate Bayesian MTC methods for specific treatments.[a]
3	Our results raised some concerns about the validity of the Bayesian MTC method for one closed loop networks, and possibly for loop patterns (but less concern). Future simulations could include additional variation of the underlying data and complexity of network patterns to determine whether these findings are consistent.
3	Our findings for one closed loop networks raise questions about the impact of a single “dangling” treatment on the results. Future simulation studies could explore this further to determine whether such “dangling” treatments should or should not be routinely included in Bayesian MTC meta-analyses.
3	For our simulations, we chose to output the probability that each drug was the best. Future research could explore whether using different outputs (e.g. mean treatment response) would either alter conclusions or uncover additional findings.

[a]For example, we could develop a simulated dataset that resemble the second-generation antidepressants or the biologic DMARDs real-world data—including setting the pre-determined truth to match our real-world findings and setting the number and sample sizes of included studies to match the real-world data. Then, we could run thousands of simulations to determine the validity of Bayesian MTC methods for a simulation that very closely matches a real-world dataset for a particular treatment and health condition.

Abbreviations: SGAs = second-generation antidepressants

Further research is needed to explore additional real-world datasets and simulated data to determine if these findings are generalizable and to better understand the accuracy and validity of Bayesian MTC methods for various scenarios. We hope this research will inform Evidence-based Practice Centers and others conducting CERs about Bayesian MTC methods and help to inform additional research and guidance for the use of these methods.

References

1. Agency for Healthcare Research and Quality. What Is Comparative Effectiveness Research. Rockville, MD. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/. Accessed May 3 2012.

2. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997 Jun; 50(6):683-91. PMID: 9250266.

3. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. Health Technol Assess. 2005 Jul; 9(26):1-134, iii-iv. PMID: 16014203.

4. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep; 7(3):177-88. PMID: 3802833.

5. Dias S, Sutton AJ, Welton NJ, et al. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. Report by the Decision Support Unit. 2011. Available at: http://www.nicedsu.org.uk/TSD3%20Heterogeneity_final%20report_docx.pdf. Accessed May 4 2012.

6. Dias S, Welton NJ, Sutton AJ, et al. A Generalised Linear Modelling Framework for Pairwise and Network Meta-analysis of Randomised Controlled Trials. Report by the Decision Support Unit. Updated August 2011. Available at: http://www.nicedsu.org.uk/TSD2%20ES%20-%2005_05_11_FINAL_updated%20Aug2011.pdf. Accessed May 4 2012.

7. Jansen JP, Crawford B, Bergman G, et al. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. Value Health. 2008 Sep-Oct; 11(5):956-64. PMID: 18489499.

8. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004 Oct 30; 23(20):3105-24. PMID: 15449338.

9. Cooper NJ, Peters J, Lai MC, et al. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? Value Health. 2011 Mar-Apr; 14(2):371-80. PMID: 21296599.

10. Mills EJ, Bansback N, Ghement I, et al. Multiple treatment comparison meta-analyses: a step forward into complexity. Clin Epidemiol. 2011; 3:193-202. PMID: 21750628.

11. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet. 2009 Feb 28; 373(9665):746-58. PMID: 19185342.

12. Li T, Puhan MA, Vedula SS, et al. Network meta-analysis-highly attractive but more methodological research is needed. BMC Med. 2011; 9:79.

13. Padwal R, Klarenbach S, Wiebe N, et al. Bariatric surgery: a systematic review and network meta-analysis of randomized trials. Obes Rev. 2011 Aug; 12(8):602-21. PMID: 21438991.

14. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med. 1996 Dec 30; 15(24):2733-49. PMID: 8981683.

15. Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value Health. 2011 Jun; 14(4):429-37. PMID: 21669367.

16. Gartlehner G, Hansen RA, Morgan LC, et al. Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression: An Update of the 2007 Comparative Effectiveness Review. (Prepared by the RTI International–University of North Carolina Evidence-based Practice Center, Contract No. 290-2007-10056-I.) AHRQ Publication No. 12-EHC012-EF. Agency for Healthcare Research and Quality Rockville, MD: December 2011. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

17. Donahue KE, Jonas DE, Hansen RA, et al. Drug Therapy for Rheumatoid Arthritis in Adults: An Update. Comparative Effectiveness Review No. 55. (Prepared by RTI-UNC Evidence-based Practice Center under Contract No. 290-02-0016-I.). Agency for Healthcare Research and Quality Rockville, MD: April 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

18. Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. Ann Intern Med. 2008 Apr 1; 148(7):544-53. PMID: 18378949.

19. Canadian Agency for Drugs and Technologies in Health (CADTH). Biological response modifier agents for adults with rheumatoid arthritis. Ottawa, ON; July 2010. Available at: http://cadth.ca/media/pdf/Biologics_for_RA_TRP_Final_Recommendations_e.pdf. Accessed May 4 2012.

20. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002 Aug 30; 21(16):2313-24. PMID: 12210616.

21. Kasper S, de Swart H, Friis Andersen H. Escitalopram in the treatment of depressed elderly patients. Am J Geriatr Psychiatry. 2005 Oct; 13(10):884-91. PMID: 16223967.

22. Roose SP, Sackeim HA, Krishnan KR, et al. Antidepressant pharmacotherapy in the treatment of depression in the very old: a randomized, placebo-controlled trial. Am J Psychiatry. 2004 Nov; 161(11):2050-9. PMID: 15514406.

23. Schneider LS, Nelson JC, Clary CM, et al. An 8-week multicenter, parallel-group, double-blind, placebo-controlled study of sertraline in elderly outpatients with major depression. Am J Psychiatry. 2003 Jul; 160(7):1277-85. PMID: 12832242.

24. Tollefson GD, Bosomworth JC, Heiligenstein JH, et al. A double-blind, placebo-controlled clinical trial of fluoxetine in geriatric patients with major depression. The Fluoxetine Collaborative Study Group. Int Psychogeriatr. 1995 Spring; 7(1):89-104. PMID: 7579025.

25. Aletaha D, Strand V, Smolen JS, et al. Treatment-related improvement in physical function varies with duration of rheumatoid arthritis: a pooled analysis of clinical trial results. Ann Rheum Dis. 2008 Feb; 67(2):238-43. PMID: 17644550.

26. Aletaha D, Ward MM. Duration of rheumatoid arthritis influences the degree of functional improvement in clinical trials. Ann Rheum Dis. 2006 Feb; 65(2):227-33. PMID: 15975967.

27. Anderson JJ, Wells G, Verhoeven AC, et al. Factors predicting response to treatment in rheumatoid arthritis: the importance of disease duration. Arthritis Rheum. 2000 Jan; 43(1):22-9. PMID: 10643696.

28. Follmann D, Elliott P, Suh I, et al. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol. 1992 Jul; 45(7):769-73. PMID: 1619456.

29. Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]: The Cochrane Collaboration; 2011.

30. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20(1):37-46.

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159-74.

32. Borenstein M, Hedges LV, Higgins JPT, et al. Introduction to Meta-Analysis (Statistics in Practice): Wiley; 2009.

33. McCullagh P, Nelder JA. Generalized Linear Models (Monographs on Statistics and Applied Probability 37). 2nd ed: Chapman & Hall; 1989.

34. Spiegelhalter DJ, Abrams KR, Myles J. Bayesian Approaches to Clinical Trials and Health-Care Evaluation (Statistics in Practice): Wiley; 2004.

35. Parmigiani G. Modeling in medical decision making: A Bayesian approach. Chichester: Wiley; 2002.

36. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. J Royal Statistical Society. 2002; 64(4):583-639.

37. Carlin BP, Louis TA. Bayesian Methods for Data Analysis (Texts in Statistical Science). 3rd ed: Chapman & Hall; 2008.

38. Sterne JAC. Meta-Analysis in Stata: An Updated Collection from the Stata Journal. first ed. StaCorpLP, TX, USA: Stata Press; 2009.

39. Morton SC, Adams JL, Suttorp MJ, et al. Meta-regression Approaches: What, Why, When, and How? Technical Review 8 (Prepared by Southern California–RAND Evidence-based Practice Center, under Contract No 290-97-0001) Rockville, MD: Agency for Healthcare Research and Quality; March 2004. AHRQ Publication No. 04-0033.

40. Dias S, Welton NJ, Sutton AJ, et al. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. last updated August 2011; 2011. Available at: http://www.nicedsu.org.uk. Accessed May 4, 2012.

41. Ades AE, Mavranezouli I, Dias S, et al. Network meta-analysis with competing risk outcomes. Value in Health. 2010; 13(8).